DOCUMENT RESUME

ED 277 689 SP 028 468

TITLE	Teacher Evaluation Study. Beginning Teacher Portfolio Study, FPMS Validity Studies, Master Teacher Score Trends, Special Instruments Norming.
INSTITUTION	University of South Florida, Tampa. Coll. of Education.
SPONS AGENCY	Florida State Dept. of Education, Tallahassee.
PUB DATE	86
NOTE	234p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC10 Plus Postage.
DESCRIPTORS	*Beginning Teachers; Elementary Secondary Education; Master Teachers; *Portfolios (Background Materials); *Professional Development; *Program Evaluation; Teacher Evaluation
IDENTIFIERS	*Florida Beginning Teacher Program

ABSTRACT

The Florida Beginning Teacher Program specifies that a portfolio be maintained for each beginning teacher, and that it include a Professional Development Plan (PDP) and formative information to justify and execute the plan. Chapter 1 of this report describes the study that was undertaken to determine whether the beginning teacher portfolios provided a complete progressive record of fully documented beginning teacher activities, assessments, and recommendations. Chapter 2 is an exploratory study of the relationships between beginning teachers' PDPs and their performance as measured by the Florida Performance Measurement System (FPMS). The third chapter sets forth a plan for providing feedback to districts, colleges, and the State Department of Education on performance of beginning teachers. Chapter 4 presents a meta-analysis of four FPMS predictive validity studies. The fifth chapter includes six analyses of data from the performance observations conducted during the 1984-85 Master Teacher Program. The sixth section of the report contains analyses of the relationship between total scores and selected frame factors on Student Services and Library/Media observations conducted in the 1984/85 Master Teacher Program. The seventh chapter descibes the development of materials for the FPMS Trainer Update Workshops. The appendices offer an annotated bibliography related to chapter 1 as well as detailed information in both narrative and tabular format on the topics of chapters 1, 4, and 7. A three-page bibliography is included. (JD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED277607

TEACHER EVALUATION STUDY

Beginning Teacher Portfolio Study
FPMS Validity Studies
Master Teacher Score Trends
Special Instruments Norming

Report for 1985 - 1986

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

D. Peterson

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

College of Education

University of South Florida

BEST COPY AVAILABLE

This project was funded under a grant from the Florida Department of Education; Planned and Coordinated by the Office of Teacher Education.

TABLE OF CONTENTS

		Page
Research/Administrative Personnel and Acknowledgments		ii
Chapter 1.	A Study of Beginning Teacher Portfolios	1
Chapter 2.	An Exploratory Study of the Relationships Between Beginning Teachers' Professional Development Programs and Their Performance as Measured by the FPMS	28
Chapter 3.	Reporting Information to School Districts and Colleges on the Performance of Beginning Teachers	46
Chapter 4.	Predictive Validity Studies	59
Section 1.	Reanalysis of Predictive Validity Data 1984-1985	61
Section 2.	Development of Guidelines for Conducting Predictive Validity Studies	72
	Outline: Guidelines for Conducting Predictive Validity Studies of Measured Teacher Performance	79
Section 3.	Macro Predictive Validity Studies: Analysis and Implications for Prototype Designs	83
Chapter 5.	Study of Master Teacher Program Score Trends	85
Chapter 6.	Special Instruments Norming Studies	128
Chapter 7.	Development of Materials for FPMS Trainer Update Workshops	140
Bibliography		143
Appendices		
Appendix A (Chapter 1)		147
Appendix B (Chapter 4)		179
Appendix C (Chapter 7)		204

Research Institute

Principal Researchers

Jean M. Borg
Jeffrey Kromrey
Theodore Micceri
Donovan Peterson
B. Othanel Smith
Melva Underbakke

Staff

Judy Crosby
Kimberly DuPree
Kellie Smith
Brian Gadson
Rebecca St. Amant

Administration

Department of Education

Betty V. Fry
James Parris
Garfield W. Wilson

Consultants

Ronald D. Adams
Patricia M. Kay
Harold E. Mitzel

ACKNOWLEDGEMENTS

Beginning Teacher Study

Reviewers

Ava Belitsky
Jean M. Borg
Mary Bullerman
James Parris
B. Othanel Smith
Melva Underbakke
Lynette Wenner

Materials Consultants

Gordon E. Eade
Lore Nielson
Debbie Samec
Christine Vacha
Anne Wilson

Coordinators

Mary Jean Davis
John John
Carol Johnson
Ron Kirves
Lore Nielson
Clarence Pilcher
James Ray
Carrie Sedinger
Mike Wilkinson
Nick Williams
Anne Wilson

Researchers

Jean M. Borg
Theodore Micceri
B. Othanel Smith

Development of Workshop Materials

Jean M. Borg
Mary Bullerman
Betty V. Fry
Diann Gordon
Carol Johnson
Nellie Laird Johnson
Pam McNab
Ronald Peake
Christine Vacha

Predictive Validity Studies

Researchers	Consultants
Jeffrey Kromrey	Judy Crosby
Theodore Micceri	Thomas Fisher
Donovan Peterson	William Katzenmeyer
B. Othanel Smith	Joseph Mazur
	Harold Mitzel
	Robert Soar

Coordinators	Observers
Miriam Davis	Belle Altice
Lynn Gold	Bill Beachum
Evelyn Lynn	Gwen Bidele
Virginia Sharpe	Sandra Blackburn
	Albert Bouie
	Jewel Dickson
	Nana Hilsenbeck
	Ed Jarvis
	Rosalie Markins
	Sandra Rowe
	Dick Schuler
	Peggy Wesley

Appreciation is expressed to all of the teachers who participated in the studies described in this report.

A project of this magnitude required the involvement of a large number of people in a variety of activities. It was impossible to credit all deserving persons but we have listed those names for which we have a record. We apologize for any omissions.

CHAPTE 1

A Study of Beginning Teacher Portfolios
General Summary

Chapter 1 describes the study that was undertaken to determine whether the Beginning Teacher Program portfolios provided a complete progressive record of fully documented beginning teacher activities, assessments, and recommendations.

A data collection instrument was developed to provide for systematic recording of the contents of a sample of portfolios representing almost all grade levels and subject areas in eleven school districts.

The primary questions dealt with in this review and analysis of portfolio data are:

1. Demographics
    a) college attended, teaching experience, grade level/subjects
    b) the nature of the support staff
2. Performance observations -- their type, timeliness, and results
3. Professional Development Plan (PDP)
    a) Is it based on domains or competencies?
    b) Which competencies or domains are identified in the PDP?
    c) Do these correspond with the information contained in the screening observation?
    d) What activities are assigned to remediate required skills?
    e) Who is responsible for assessment of these training activities?
    f) What techniques are used to assess teacher remediation?
4. Documentation of Competencies
    a) Which have supporting documentation and which are merely stated as having been documented?
    b) What types of documentation are used for particular competencies.
    c) Do the types of documentation correspond with ideal types?
5. Conferences
    a) When, and how frequently, do these occur?
    b) What types of conferences occur?
    c) Who attends these conferences?

The data analyses of the portfolio sample point up the variation in the amount of information in the portfolios. The poverty of evidence from the screening procedure leaves one wondering just how much was done to determine what specific behavior the BT needed to remediate. Without adequate diagnosis what follows may or may not be meaningful for the teacher, and also, for the determination of what is good professional assistance. While ninety four percent of the portfolios contained at least one screening observation, only 70 percent of the valid cases met the State Board Rule that specifies diagnostic (screening) observation be administered within the first 45 working days of employment. Furthermore, although deficiencies were identified as indicator-items on the screening instrument, in almost 52 percent of the portfolios the indicators were not used in the selection and organization of remedial activities. Given this condition, focus of a corrective program would be so general, or so misplaced, as to induce little hope of change in pre-post assessment of specific behavior as a result of that program.

Much of the missing information can be attributed to the practice of checking off competency demonstration on a list of types of documentation. Therefore, since the summative observation was commonly used to document all of the competencies, although it covers only nine of the twenty-four, almost all portfolios did contain some material evidence of documentation, albiet suspect evidence; only six percent had none. Without more evidence it would be difficult to weigh the effectiveness of the methods used to assess the successful demonstration of the individual competency, i.e. it is hard to see how a work sample supports number 1 of the essential competencies -- oral communication. If support deviates from the suggested types recommended in the BTP that evidence should surely be presented and evaluated.

In the earlier studies there were reports that BTs were not getting the help they thought they should have. The portfolio provided a record of the amount of BT activity scheduled with the support staff. For example, an average of 3.6 conferences per BT in the 104 portfolios reporting conferences, and over half of the conferences were pre or post observation (53 percent). Another indication of assistance was the amount of interaction with the support staff that was recommended in the professional development plan. In some 17 percent of the activities assigned to remediate problems, BTs were to receive coaching. In addition, the records do not show the relationship between the prescribed workshops (nine percent) and the diagnosed specific behavior to be

remediated. The workshop content could have been applicable to a few or to all BTs, or even all teachers. By far the largest category, "other /primarily peer or supervisor recommendations" (40 percent), included such directions as to practice giving more praise or more academic specific praise, smile more, or take a course on classroom discipline, be more positive, organize your materials before class, etc. Exhorting a BT to do a specific behavior does not usually change the performance nor provide assistance in developing that skill. The identification of the specific behaviors to be remediated was a necessary part of the professional growth program, providing the diagnosis was done objectively by formative instruments and/or other techniques for identifying inadequacies of teacher performance, but the identification fell short of providing the activities to assist the BT in developing the skills. (It's hard to smile more when the kids are driving you crazy -- and smiling might make matters worse.)

Finally, to summarize there is not much reported on the BTs' college career (almost 40 percent did not list the institution from which the BT was graduated). Limited information on other demographic items (BTs' teaching experience, peer teachers' subject area, other educators' job or work location, etc.) prohibit much analysis of relationships among demographic variables, and suggest finding alternative ways of compiling and matching this information in the future.

It was learned from the screening observations that not all schools comply with the rule to observe BTs during the first 45 working days (70 percent reported screening observations during September and October).

The mode of identifying deficiencies in the areas of the teacher's performance was not well documented. In 28 percent of the cases the PDP was developed on the competencies to be documented without evidence of diagnosis. The competencies most frequently identified for remediation were (1) oral communication, (6) student growth, (7) diagnosing entry level skills, (9) sequencing objectives, (10) selection of materials, (11) sequencing activities, (12) establishing rapport, (16) formulation of student behavioral standards, and (17) formulation of behavioral management techniques. Domains 2.0 (management of student conduct) and 3.0 (instructional organization and development) were the most frequently named domains in the PDPs. The most commonly coded ineffective indicators on the screening observations were all from Domains 2.0 and 3.0 with the exception of ineffective treatment of concepts (16 percent) from Domain 4.0.

The sample of methods used to assess completion of remedial activities (28 cases) and person responsible for assessment (30 cases) was small but showed that observation by the BLA was the most common procedure.

Except for the cognitive exam, every competency was evidenced by every type of documentation. This leads one to wonder which has occurred: evaluation or just blanket competence checking off. Discrepancies can be easily seen by comparing Table 17 with the BTP recommended documentation (Appendix A).

One would expect first conferences to correspond with the screening observations, preferably as pre observation conferences to explain the procedure, etc. However, the data show first conferences occurring throughout the year, but generally before December. The lack of information could be the result of poor record keeping or the problem of contact that was reported in earlier studies.

The value of fully documented files for program accountability within the school district seems defensible. To assist school districts in developing a format that provides that accountability for the professional growth of BTs would be the benefit of this study. Guidelines for worthwhile data collection and local monitoring of the BTP should be developed from the knowledge compiled herein.

CHAPTER 1

A Study of Beginning Teacher Portfolios

The Florida State Board of Education Rule (Chapter 6A-5.075) for the Beginning Teacher Program (BTP) specifies that a portfolio be maintained for each beginning teacher (BT), and that the portfolio include, but not be limited to, a professional development plan and formative information to justify and execute the plan.

This is a study of a sample of portfolios. Its purpose is it to a) ascertain the content of the portfolios, b) evaluate the portfolios, and c) determine the relation between the quality of portfolios and changes in the classroom performance of BTs. This chapter deals with the question of portfolio content. Treatment of the question of evaluation and the bearing of portfolio quality on teacher performance is reserved for the next chapter.

It should be emphasized that this study of portfolios and the effect of the quality of portfolios upon BT performance is a continuation of a monitory process, which began as a questionnaire review of the program in its first year of operation, and continued in 1985 as an interview study of a sample of 143 BTs in ten school districts.

The primary questions dealt with in this review and analysis of portfolio data are:

1. Demographics
    a) college attended, teaching experience, grade level/subjects
    b) the nature of the support staff
2. Performance observations -- their type, timeliness, and results
3. Professional Development Plan (PDP)
    a) Is it based on domains or competencies?
    b) Which competencies or domains are identified in the PDP?
    c) Do these correspond with the information contained in the screening observation?
    d) What activities are assigned to remediate required skills?
    e) Who is responsible for assessment of these training activities?
    f) What techniques are used to assess teacher remediation?

4. Documentation of Competencies
    a) Which have supporting documentation and which are merely stated as having been documented?
    b) What types of documentation are used for particular competencies.
    c) Do the types of documentation correspond with ideal types?
5. Conferences
    a) When, and how frequently, do these occur?
    b) What types of conferences occur?
    c) Who attends these conferences?

Data-Collection Procedure

This study began by identifying several seemingly important types of possible portfolio information, creating an instrument to collect data on these types, and training a group of persons to use the instrument in a study of a sample of BT portfolios selected from 11 school districts. Specifically the following procedure was followed:

In order to standardize the analysis of portfolios an instrument for data recording was constructed. The following items were built into the instrument: 1) identification of BTs, 2) identification of support staff, 3) observations of BTs (dates, types of instruments), 4) professional development plan, 5) documentation of competencies, and 6) record of conferences.

In its initial form, the review instrument was tried out by six potential reviewers, using portfolios provided by a school district. The inadequacies of the instrument, learned from the try out, were corrected and the instrument then revised for the portfolio reviews.

A sample of 133 BTs was selected from 11 school districts as follows: Alachua, Bay, Broward, Citrus, Escambia, Hardee, Hillsborough, Jackson, Madison, Pinellas, and Polk. Ten of these districts were the same as the sample used in the 1984-85 Interview Study of BTs, Broward district being the exception. This sample was originally selected to represent small, medium, and large districts (by number of students), rural and urban areas, and state geographic regions -- south, west central, crown, and panhandle. The addition of Broward expanded the sample to represent an urban area of south Florida. Participation in the study was voluntary.

The sample included only full-year beginning teachers who began their work in the fall of 1984. They represented almost all grade levels and subject areas, and where possible no more than one teacher from a school, grade, and subject.

The school districts were contacted between January 7 and 10 and arrangements made for the reviews, which were completed between January 18 and February 7.

The Florida State Department of Education arranged with the districts for the portfolios to be collected and located at a central meeting place. A group of two to four reviewers visited each district. This distribution of reviewers made it feasible for each reviewer to process from three to six portfolios, thus reducing the effect of fatigue upon the observation and recording of data. Exceptions to this procedure were Alachua, Escambia, and Broward Districts where there was only one reviewer.

Demographic Findings

Some information was not generally available in the portfolios. In particular, information regarding a teacher's college career (other than the name of the institution) was almost nonexistent. It can be seen from Table 1 that the identification of the institution attended by the BT was available for 81 cases or 61 percent of the portfolios. Of these 81 cases, 32 cases or almost 40 percent, reported as having been graduated from out-of-state colleges. No Florida college or university was represented by more than 12 cases or 15 percent of the valid responses. Almost 40 percent of the sample, or 52 out of 133 portfolios, failed to identify the institution from which the BT was graduated.

TABLE 1

Institutions Represented in Survey

College	N Graduated	Pct
FAMU	2	1.5
FLA SOUTHERN	5	3.8
FSU	11	8.3
FAU	4	3.0
UF	7	5.3
USF	7	5.3
UWF	12	9.0
OUT OF STATE	32	24.1
Not Given	52	39.1
TOTAL	133	100.0
valid cases 81		missing cases 52

The great majority of portfolios lacked information regarding a teacher's experience. As Table 2 shows, 46 of the 133 portfolios contained no record of the BT's experience. Of those

TABLE 2

Beginning Teacher's Experience

Years Experience	N Cases	Pct
0	68	77.2
1	10	11.3
2	4	4.5
4	3	3.4
5	1	1.1
6	1	1.1
TOTAL	87	100.0
valid cases 87		missing cases 46

87 cases containing this information, 77 percent indicated no prior experience.

A wide range of grade levels and subject areas were represented in the sample, as shown in Tables 3 and 4.

TABLE 3

Grade Level of Beginning Teachers

Grade Level	N Cases	Pct
Elementary	37	27.8
Middle	38	28.6
Secondary	38	28.6
Not given	20	15.0
TOTAL	133	100.0
valid cases 113		missing cases 20

No grade level is indicated in 20 cases or 15 percent of the sample, and 16 cases or 12 percent of the sample are not identified as to teaching field.

TABLE 4

Subject Area of Beginning Teachers

Subject Area	N of Teachers	Pct
Language arts	14	10.5
Foreign language	7	5.3
Social sciences/history	13	9.8
Mathematics	13	9.8
Science	14	10.5
PE or ROTC	7	5.3
Business education/DCT	1	0.8
Industrial education	2	1.5
Vocational education	7	5.3
Arts	3	2.3
Music	3	2.3
Exceptional child ed	13	9.8
Elementary education	20	15.0
Not given	16	12.1
TOTAL	133	100.0
valid cases 117		missing cases 16

Support Staff

The support staff, according to State Board of Education Rule, must consist of at least the Building Level Administrator (BLA), peer teacher, and other professional educator. The rule further specifies that the peer teacher "hold a valid regular certificate, and teach at the same level and in the same subject area or the same service area as the BT." The other professional educator shall be a professionally trained and experienced individual. Among these educators the following are mentioned: "teacher education center directors, staff development specialists, curriculum directors, instructional supervisors or specialists, college or university teacher educators."

By summing the percent of principals and assistant principals in Table 5, it is clear that in 98 percent of the reported cases the BLA on the support staff was either a principal or assistant principal. As can be seen from inspection of the same Table, information as to BLA was lacking for 21 percent of the total cases.

TABLE 5

Building Level Administrator Position

Position	N Cases	Pct Reported
Principal	92	87.6
Assistant Principal	11	10.5
Other	2	1.9
Not given	28	----
TOTALS	133	
valid cases 105		missing cases 28

Information regarding the job or work location of Other Educators on the support staff was generally unavailable in the portfolios. However, for the portfolios containing information about Other Educators (N=70), some 56 percent came from district level offices, while the remaining 44 percent came from school level offices (Table 6).

The support staff is a significant feature of the PDP. Unless there is a close working relationship among members of the support staff, especially between the peer teacher and the BT, the assistance received by the BT is likely to be weak with respect to problems of instruction in the BT's field.

TABLE 6

Position Level of Other Educator

Location of Other Educator	N Cases	Pct
District level	39	29.3
School level	31	23.3
Not given	63	47.4
TOTAL	133	100.0
valid cases 70		missing cases 63

An attempt was made to match the beginning teacher's subject area with that of the peer teacher. Since only limited information was available, comparisons were possible for only 34 cases. In only 20 of the 34 cases, or 59 percent, did the peer teacher area match directly with BT's subject area. Seven cases fell into the category: other; and for the remaining seven cases, the match between peer specialization and beginning teacher subject area is less obviously appropriate:

Figure 1

Peer Teacher	Beginning Teacher
Exceptional Child	Language Arts
Industrial Education	Mathematics
Language Arts	Exceptional Child
Home Economics	Exceptional Child
Language Arts	Foreign Language
Physical Education	Vocational
Elementary Education	Social Sciences

Screening Observations

The State Board Rule specifies that the diagnostic (screening) observation be administered within the first 45 working days of employment. It further specifies that information gained from the screening observation be used in planning the PDP of the teacher. In addition, a summative evaluation is specified to verify successful performance of the essential competencies. To what extent were these requirements complied with in the sample of portfolios?

TABLE 7

Number of Summative Observations in Portfolio

Number in Portfolio	N Cases	Pct
0	8	6.0
1	8	6.0
2	82	61.7
3	27	20.3
4	6	4.5
5	1	0.8
7	1	0.8
TOTAL	133	100.0

valid cases 133 missing cases 0

Some 94 percent of all portfolios contained at least one summative observation, while 88 percent contained at least two summative observations made during the program (Table 7).

Table 8 shows that the time of the screening observations was not given in 12 percent, or 16 cases of the sample of screening observations, leaving a total of 117 valid cases. Of these 117

TABLE 8

Time of Screening and Final Observations

Month	Screening		Final	
	N Cases	Pct	N Cases	Pct
September	14	10.5	1	0.8
October	79	59.4	1	0.8
November	9	6.8	1	0.8
December	3	2.3	2	1.5
January	4	3.0	7	5.3
February	2	1.5	15	11.3
March	2	1.5	14	10.5
April	0	0.0	0	0.0
May	4	3.0	62	46.6
Not given	16	12.0	28	21.1
TOTALS	133	100.0	133	100.0

valid cases 117 valid cases 105

cases 14 or approximately 11 percent were made during September and 79 cases or about 59 percent in October. Assuming that most Florida schools open in mid August, it is reasonable to

assume that 70 percent of the valid cases were observed within the first 45 working days. Twenty-four cases or 18 percent were observed between November and May.

Final observations were generally conducted late in the year, with some 46 percent occurring in April or May (Table 8).

Identification of Deficiencies

One of the primary objectives of the BTP is to improve the classroom performance of the beginner. This objective requires that the deficiencies in the BT's performance be identified and that a corrective program be developed, administered, and changes in the teacher's behavior measured and evaluated.

Two procedures are provided by the BTP for identifying teacher deficiencies. The first involves the use of an observation instrument, referred to as a screening instrument, which enables a trained observer to identify the problem areas, such as management of student conduct, organization and conduct of instruction, and teaching basic elements of knowledge, in which the teacher is experiencing difficulties.

The second procedure provides for formative observation instruments with which to pin-point the specific teacher behaviors that require remediation. For example, the screening observation may indicate that the teacher is experiencing trouble with student misbehavior, but the specific obstacle may not be indicated. The specific difficulty, for example, fragmenting group movement, can be identified by observing the teacher's performance more closely with an appropriate formative instrument.

These two procedures, however, do not cover all of the original 24 competencies required by the State for certification. Nine of 24 are included in the screening instrument and three of the remainder are covered in the formative instruments, making a total of 12. The twelve remaining competencies are assessed by a medley of recommended techniques. Among these are successful completion of a workshop, learning package, etc., teacher prepared learning center, sample of teacher made tests, record of referral, or performance on Florida Teacher Certification Examination.

Turning to the deficiencies identified by observations made with the screening (first) instrument, the indicators of ineffective, rather than effective, performance were analyzed since

they are associated with the teacher's problems (Table 9). Indicators were delays (31 percent), use of non-academic questions (29 percent), use of general praise (26 percent), disorganization of

TABLE 9

Ineffective Indicators Noted on Screening Observation

All Cases

Indicators	N Responses	Pct Responses	Pct Cases
Delays	31	17.2	31.0
Disorganization of materials	23	12.8	23.0
Unrelated talk/activity	18	10.0	18.0
Multiple questions	5	2.8	5.0
Non-academic questions	29	16.1	29.0
Harsh response or ignores	3	1.7	3.0
General praise	26	14.4	26.0
No practice	2	1.1	2.0
No homework/feedback	3	1.7	3.0
Remains at desk	13	7.2	13.0
Treats concepts poorly	16	8.9	16.0
Inadequate cause/effect	5	2.8	5.0
No criteria for judgment	1	0.6	1.0
Vague/scrambled discourse	1	0.6	1.0
Frowns/deadpan/lethargic	1	0.6	1.0
Loses momentum	3	1.7	3.0
Total Responses	180	100.0	180.0
valid cases 100		missing cases 33	

materials (23 percent), allowed unrelated talk or activity (18 percent), ineffective treatment of concepts (16 percent), and remained at desk (13 percent). The several Tables 9 - 13 were based on different numbers of total cases ranging from 39 to 100. The numbers and percentages among these tables will therefore not correspond precisely.

Evidence of the use of formative instruments to diagnose specific deficiencies in the areas of trouble identified by the screening observations instrument was not found in the portfolios. As will be seen later, however, formative instruments were used to document proficiency in certain competencies. Furthermore, the sample of portfolios contained no evidence that the medley of techniques referred to earlier were employed in the process of identifying inadequacies of teacher performance. Since the screening instrument embraces only nine of the State's 1984-85 prescribed

competencies, the lack of evidence that other procedures and techniques were used to identify deficiencies raises a puzzling question: How were the teacher's ineffective behaviors identified in the remaining fifteen competencies?

Professional Development Program

Although deficiencies were identified as indicator-items of the screening instrument, the indicators were not used in the selection and organization of remedial activities. Instead the PDPs were centered in the domains of the knowledge base of the FPMS and on the competencies, although the mode of identifying deficiencies of the latter was not available in the portfolios. It can be seen from Table 10 that 69 portfolios, or almost 52 percent of the PDPs were organized by Domains rather than by indicators. For example, a teacher who fails to give either appropriate or

TABLE 10

PDP Organized by Competencies or Domains?

Type of Organization	N Cases	Pct
Competencies	30	22.6
Domains	69	51.9
Competencies and Domains	8	6.0
Not given	26	19.5
TOTAL	133	100.0
valid cases 107		missing cases 26

ample examples of a concept may be directed to study Domain 4.0 of the knowledge base of the FPMS instead of study of, and practice in, the selection and testing of examples.

Table 10 shows that PDPs were included in about 80 percent of the portfolios indicating some 20 percent did not receive help. Of the portfolios containing PDPs, some 72 percent were documented either by domain or a combination of domain and competency, while only 28 percent were documented on a strictly competency basis. Table 11 indicates that the competencies most frequently identified in the PDP included those of (1) oral communication, (6) student growth, (7) diagnosing entry level skills, (9) sequencing objectives, (10) selection of materials,

(11) sequencing activities, (12) establishing rapport, (16) formulation of student behavioral standards, and (17) formulation of behavioral management techniques. For those portfolios documented primarily using domains, the Domains 2.0 (Management of Student Conduct) and 3.0 (Instructional Organization and Development) received the most attention in the PDPs (Table 12).

TABLE 11

Which Competencies Were Worked on in PDPs

Competency Number	Competency	N Response	Pct Response	Pct Cases
1	Oral communication	20	4.9	51.3
2	Written communication	13	3.2	33.3
3	Listening comprehension	14	3.4	35.9
4	Professional readings	15	3.7	38.5
5	Mathematics skills	11	2.7	28.2
6/19	Encourage student growth	19	4.7	48.7
7	Diagnose entry level skills	23	5.7	59.0
8	Identifies goals	18	4.4	46.2
9	Sequences objectives	20	4.9	51.3
10	Selects materials	28	6.9	71.8
11	Sequences activities	24	5.9	61.5
12	Establishes rapport	21	5.2	53.8
13	Directs	16	3.9	41.0
14	Proper testing procedures	18	4.4	46.2
15	Establishes routines	15	3.7	38.5
16	Sets behavioral standards	20	4.9	51.3
17	Uses behavioral mnge tchns	23	5.7	59.0
18	Keeps records	16	3.9	41.0
20	Student dignity/worth	6	1.5	15.4
21	Student self concept	19	4.7	48.7
22	Encourages peer interaction	19	4.7	48.7
23	Values	14	3.4	35.9
24	Deals with exceptional students	15	3.7	38.5
TOTAL		407	100.0	1043.6

valid cases 39 missing cases 94

TABLE 12

Domains Worked on in PDP

Domain Number	Domain Titles	Referrals for Remediation N	Pct.
	Planning	36	18.9
2	Management of Student Conduct	52	27.4
3	Instructional Organization and Development	46	24.2
4	Presentation of Subject Matter	12	6.3
5	Communication: Verbal and Nonverbal	26	13.7
6	Testing: Student Preparation, Administration And Feedback	18	9.5
Total		190	100.0

valid cases 72 missing cases 5

TABLE 13

Activities Designed to Remediate Problems

Activity	N Referrals	Pct Referrals	Pct Cases
Professional readings	22	18.8	73.3
Observes other teachers	11	9.4	36.7
Studies videotapes	2	1.7	6.7
Self instruction	5	4.3	16.7
Receives coaching	20	17.1	66.7
Other /primarily peer or supervisor recommendation	47	40.2	156.7*
Total Responses	117	100.0	.0

valid cases 30 missing cases 103

* Some of these have more than one activity scheduled per case, since more than one competency or domain may be remediated, often using the same activity.

Remedial Activities

Having identified areas of teacher performance, requiring remediation, activities were assigned in the PDP to remedy these deficiencies. Table 13 shows the most commonly assigned

remediation activities to be: professional readings (19 percent of the total number of activity referrals), coaching (17 percent), and those falling into a general category of other (40 percent).

Table 14 shows that several methods were used to assess the completion of the activities, and the most commonly used was that of observation (49 percent of cases).

TABLE 14

Methods Used to Assess Completion of Remedial Activities

Assessment Methods	N Responses	Pct Responses	Pct Cases
Interview	22	20.0	78.6
Observation	54	49.1	192.9
Work samples	34	30.9	121.4
Total Responses	110	100.0	392.9
valid cases 28		missing cases 105	

The responsibility for assessing assigned activities fell primarily to the BLA (47 percent of activities), with the peer teacher individually responsible for 29.7 percent and the support staff as a whole responsible for 17 percent of the cases. The other educator only bore this responsibility for six percent of the assigned activities (Table 15).

TABLE 15

Personnel Responsible for Assessment of Remedial Activities

Support Staff Member	N Activities	Pct Responses	Pct Cases
BLA	133	47.0	443.3*
Peer teacher	84	29.7	280.0
Other educator	18	6.4	60.0
All support staff	48	17.0	160.0
Total Responses	283	100.0	943.3
valid cases 30		missing cases 103	

* One individual may be responsible for as many as 23 activities per case.

Documentation of Competencies

Following the assessment of remedial activities, the rule requires documentation of the State competencies for the BT. At least some physical evidence of documentation for a particular competency was present in 77 percent of the portfolios reviewed. As Table 16 shows, those competencies having the least reported documentation were: (3) listening comprehension - 53 percent, (23) facilitating values clarification - 49 percent, and (24) recognizing the needs of exceptional students - 45 percent.

Of the 3589[1] documentations of competence in 118 of the 133 portfolios, the most common techniques were: formative observations - 28 percent, work samples - 20.5 percent, and screening observations - 19 percent. By far the least common technique was that of cognitive exams - three percent. Several competencies were documented more frequently than most by cognitive exam, particularly: (5) mathematics concepts - 13 percent, and (21) assisting students with positive self-concept development - 10 percent of the cases. Those competencies least frequently using screening observations were (2) written communication - 17 percent, (4) comprehension of professional reading - 16 percent, (10) selecting materials - 12 percent and (14) the construction of tests - 14 percent. Table 17 shows the percentage of cases for which various types of documentation are used for each competency.

---

[1] This is the combined total of 23 competencies, some with multiple documentation, across 118 cases.

TABLE 16

Documentation - Percentage Claimed, and Present in Portfolio

Competency Number	Competency	Percent of Total Cases: Claiming Documentation (n=133) %	Percent of Total Cases: Showing Documentation in Portfolio (n=133) %
1	Oral communication	78	69
2	Written communication	77	72
3	Listening comprehension	57	53
4	Professional readings	67	62
5	Mathematics skills	65	61
6/19	Encourage student growth	74	68
7	Diagnose entry level skill	69	67
8	Identifies goals	71	69
9	Sequences objectives	77	76
10	Selects materials	69	68
11	Sequences activities	64	63
12	Establishes rapport	69	65
13	Directs	67	66
14	Proper testing procedures	77	77
15	Establishes routines	72	70
16	Sets behavior standards	73	72
17	Uses behavioral mnge tchn	67	64
18	Keeps records	59	57
20	Student dignity/worth	58	57
21	Student self concept	54	52
22	Encourages peer interactions	56	54
23	Values	51	49
24	Deals with exceptional students	48	45

TABLE 17

Breakdown of Documented Support By Competency

	Type of Documentation						
	COG EXAM %	WORK SAMPLE %	SUMATV OBS %	FORM OBS %	INTR VIEW %	EX-HIBIT %	COMB WORK EXHB %
(1) Oral Communication	3	17	38	54	24	9	14
(2) Written Communication	6	55	17	34	6	22	10
(3) Listening Comprehns	5	16	30	38	29	13	9
(4) Professional Rdgs	8	38	16	31	21	19	15
(5) Math Skills	13	37	27	33	14	22	3
(6/19) Encourage Stu Grwth	3	29	32	35	18	18	15
(7) Diagnose Entry Skill	6	36	26	32	22	19	11
(8) Identify Goals	6	44	17	30	21	18	14
(9) Sequence Objectives	3	44	20	30	12	18	16
(10) Select Materials	4	47	12	31	10	19	12
(11) Sequence Activities	8	33	23	33	17	19	15
(12) Establishes Rapport	=	13	37	57	15	18	10
(13) Directs	=	15	43	62	9	13	5
(14) Adequate Testing	1	48	14	38	8	25	17
(15) Establishes Routines	1	28	37	48	15	13	8
(16) Sets Beh Standards	3	31	36	46	9	16	11
(17) Uses Beh Mge Techq	4	22	31	46	22	11	11
(18) Maintains Records	3	44	21	33	13	30	13
(20) Student Dignity/Wrth	4	18	37	49	19	11	7
(21) Student Self-concept	10	17	37	51	17	12	6
(22) Peer Interactions	7	15	36	56	16	16	3
(23) Values	6	15	36	51	17	16	6
(24) Exceptional Students	7	28	26	42	13	15	14

* This table is based on multiple possible sources of documentation for approximately 110 of the 113 total cases (it varies from competency to competency). There is between five and ten percent error in the percentages given above. Each percent represents the percent of cases for which this form of documentation was used for a particular competency.

Conferences

An average of 3.6 conferences was reported in the 104 portfolios reporting conferences. Table 18 shows that most conferences are reported as pre observation (29 percent), with only about 10 percent reported as other than either pre observation or follow-up.

TABLE 18

Types of Conferences Held

District	N Confs	Pct
Pre observation	104	29.0
Post observation	91	24.0
Follow-up	78	20.6
Final assessment	64	16.9
Other	36	9.5
Total	374	100.0
valid cases 100	missing cases 29	

Table 19 shows that first conferences generally occurred prior to December, while final assessments, generally occurred late in the year, with fully 55 percent of those documented (n=60) occurring after February.[2]

Table 20 suggests that pre conference meetings are attended by peer teachers and other educators more often than by BLAs, post observation conferences by all three of the support staff, with peer teachers most common, and final assessments primarily by the BLA if only one support staff member is present. Other types of conferences frequently include the other educator. overall, when only one member of the support staff is present at a conference it tends to be the BLA (44 percent of the time), with the peer teacher (30 percent) and the other educator (26 percent) present about an equivalent percentage of the time.[3]

---

[2] A large number of portfolios contained no information regarding either first (pre-observation) conferences or later ones.

[3] These percentages are computed on the 123 cases of conferencing attended by only one member of the support staff(Table 21).

TABLE 19

Timeliness of Conferencing

Month	First Conference N Cases	Pct	Last Conference N Cases	Pct
September	11	8.3	1	0.
October	45	33.8	2	1.
November	14	10.5	2	1.
December	8	6.0	4	3.
January	3	2.3	8	6.
February	5	3.8	9	6.
March	5	3.8	13	9.
April	2	1.5	20	15.
May	3	2.3	1	0.
Not given	37	27.8	73	55.
Total	133	100.0	133	100.0
valid cases 96			valid cases 60	

TABLE 20

Conferencing and Support Staff Attendance

Type of Conference	Who Attends Which Type of Conference				
	BLA N*	PEER TCHR N	OTHER EDUC N	ALL THREE N	TOTAL N
Pre Observation Conferences	4	12	10	13	39
Post Observation or Follow-Up	11	21	12	20	64
Final Assessment	10	1	4	8	23
Other	5	3	6	2	16

* This table results from cross-tabulations in which a missing case for one variable is missing for all, therefore, the numbers will not match up with those produced by either simple or multiple frequency distributions.

TABLE 21

Who Attends Conferences

Support Staff Member	N* Confs	Pct Confs
BLA	54	32.3
Peer teacher	37	22.2
Other educator	32	19.2
All support staff	44	26.3
Total	167	100.0

valid cases 86 missing cases 47

* More than a single conference occurs per case.

Relationship Between Screening Observations and PDPs

Because of the nature of the information contained in portfolio files, this relationship must be investigated in a somewhat indirect fashion. Several factors must first be considered.

1. It was not possible to interpret specific effective indicators or domains from a screening observation for purposes of remediation, since interpreting these types of behaviors generally require lesson specific knowledge. On the other hand, ineffective behaviors almost always require remediation. For this reason, the comparison was made on the basis of ineffective behaviors coded on the screening observation.
2. Due to the small number of cases (n = 23) documented by competency, the even smaller number of cases containing ineffective indicators, and the large number of competencies (n = 23), it proved impossible to obtain useful information when attempting to map ineffective indicators to PDP competency assignments. For this reason, analysis was limited to those cases documented by domains and containing ineffective indicators on the screening observation (n=46).
3. It was impossible to make a direct one-to-one mapping of ineffective indicators to PDP domain assignments, because a specific portfolio contained more than

one ineffective indicator and more than one domain assignment. Thus, only a very rough indication of the relationship between observations and PDP remediations was achieved.

The mapping of ineffective indicators to PDP assignments is shown in Table 22. This table represents a synthesis of 39 separate Tables designed to increase comprehension. For this reason some of the numbers appear not to match properly. In cross-tabulation procedures, if a case is missing for one variable but not the other, the case becomes a missing case for both. Thus, raw frequency tables shown in other parts of this report will contain more cases than the following tables. In several instances, more than 100 percent of cases are represented in the PDP domain assignments. This occurs because some portfolios contain PDP domain assignments relevant to a specific indicator for other reasons than that indicator's presence on the screening observation (for example, Domain 3 assignments may relate to several different indicators).

If a BT is coded as exhibiting one or more ineffective indicators on the screening instrument, the support staff may assign the teacher to study (read the research, examine examples, develop examples, etc.) the relevant domain(s). There are two types of such assignments. In the first type, the BT is asked, it is assumed, to study the domain to which the ineffective indicator(s) belong. For example, if a BT is having trouble in concept teaching, s/he may be asked to study Domain 4 to which the indicator "treats concepts poorly" belongs. In the second type, the BT may be assigned to study one or more domains whose research and examples are related to the ineffective indicator but to which the indicator does not belong. For example, "treats concepts poorly" belongs to Domain 4, but the BT may learn about how to teach concepts from a study of Domain 1, Planning, as well as from Domain 4, and may thus be assigned to study that Domain also.

Table 22 shows that there are 9 BT who exhibited the ineffective indicator "unrelated talk/ activity." Sixty-seven percent, or six cases, were assigned to Domain 3. Three of the nine cases were assigned to Domain 2, for this domain also was deemed to be relevant to the indicator. In addition, a number of the cases were assigned not only to Domain 3 but also to Domain 2, making altogether sixty-seven percent assigned to the latter Domain.

For all ineffective indicators, except "general praise" (33 percent) and "No homework/ feedback" (50 percent), two-thirds or more showed identification with the related domain. Table

22 suggests that generally speaking domain assignment was appropriate for those behaviors identified in the screening observation. The only questionable assignments were where the relevant Domain was not included in the assignment of domains, that is, "general praise," "unrelated talk/activity," "non-academic questions," and "no homework/feedback." The general praise indicator was not always ineffective (grades k-3), and was interpreted by some support-staff as occasionally appropriate at higher grades also.

TABLE 22

Comparison of Ineffective Indicators with Relevant Domains

Percentage Basis

		Total Documented Cases (n=46) Including at Least One Ineffective Indicator on Screening Observation					
			Assigned Domain in PDP				
Domain-Source of Indicator	Ineffective Indicator	Cases Having Indicator N	1 %	2 %	3 %	4 %	5 %
3	Delays	11	91		100		
3	Disorganized	9	67		100		
3	Unrelated talk/activity	9		67	67		
3	Multiple questions	1			100		
3	Non/academic questions	13			85		38
3	General praise	15			33		40
3	No homework/feedback	2			50		
3	Remains at desk	4		100			
4	Treats concepts poorly	5	100		100	80	
4	Inadequate cause/effect	1	100		100	100	

Discussion

Is the BT portfolio a viable part of the BTP? Does it provide the measure of data that would help determine what works -- what is successful as professional assistance to BTs during that first year in the schools?

Ideally, the portfolio should provide a complete progressive record of fully documented BT activities, assessments, and recommendations. To investigate the adequacy of the portfolio an

exploratory study was undertaken. It began by developing an instrument to standardize the collection of information by trained reviewers from a sample of BT portfolios representing almost all grade levels and subject areas in 11 school districts. The data were classified into five areas, namely: demographics, performance observations, professional development plans, documentation of competencies, and conferences.

The findings of this study indicate not only the need for a more systematic plan of record keeping, but also for school districts to be more diligent in collecting and using the data called for by State Board Rule. (For proposed changes in State Board Rule see Appendix A.)

CHAPTER 2

An Exploratory Study of the Relationships Between Beginning Teachers' Professional Development Programs and their Performance as Measured by the FPMS

General Summary

Florida State Board of Education Rule requires that a professional development plan, based on formative information from screening observations, be formulated for each beginning teacher (BT). Do these plans make a difference in BT performance? More specifically, is there a relationship between scores resulting from classroom observations -- the total scores and subscores -- and 1) the number of remedial activities, and 2) the value of the activities? To answer this question we investigated the relationship between the professional development plans and the scores of BTs on pre and post FPMS classroom observations.

Those elements of the development plans most likely to impact teacher behavior are the activities in which teachers participate. Examples of these activities as determined by inspection of BT portfolios, referred to here as remedial activities are workshops, reading assigned references, studying the FPMS Domains, and being observed by the principal or peer teachers.

Overall, the effect of the various combinations of activities reported in the portfolio sample can be summarized as follows:

1. An effect scale was constructed, based upon the findings of a meta-analysis study of 91 research studies of the effectiveness of inservice procedures and findings. When the remedial activities, reported in the portfolios, are ranked by this scale, 45 percent of the portfolios have activities of an average value of 3 or above on a scale of zero to four.
2. The effective behavior for this sample of BTs does not appear to be changed by observation, remedial activities, or conferences.
3. The ineffective behavior appears to be improved by the number of formative observations and possibly by the total number of observations (both summative and formative) per teacher.

4. There is apparently no consistent relationship between the number of formative observations conducted and the number of remedial activities assigned per teacher.
5. In this sample, Domain 4.0 is the only domain where the data suggest that the teacher's behavior may be influenced by the number of formative observations, although the relationship only approaches significance. The number of activities engaged per problem seems to positively influence teacher behavior in Domain 4.0. No other domain appears to be significantly influenced by remedial activity in this study.
6. The BT performance in Domain 4.0 appears to be improved by the total number of activities in the PDP, the total value of the multi activities in the PDP, and the mean value of the multi activities.
7. Effective scores may be negatively influenced by the mean value of single activities.

CHAPTER 2

An Exploratory Study of the Relationships Between Beginning Teachers' Professional Development Programs and their Performance as Measured by the FPMS

Florida State Board of Education Rule requires that a professional development plan, based on formative information from screening observations, be formulated for each beginning teacher (BT). The extent to which this requirement has been satisfied by our sample of schools was reported in the preceding chapter. To be explored in this chapter is the question of whether or not professional development plans make a difference in the behavior of BTs. More specifically, is there a relationship between scores resulting from classroom observations -- the total scores and subscores -- and 1) the number of remedial activities, and 2) the value of the activities? To answer this question we investigated the relationship between the professional development plans and the scores of BTs on pre and post FPMS classroom observations.

Elements of development plans most likely to impact teacher behavior are the activities in which teachers participate. Examples of these activities as determined by inspection of BT portfolios, referred to here as remedial activities, are workshops, reading assigned references, studying the FPMS Domains, and being observed by the principal or peer teachers.

There are at least two approaches to the problem of exploring the influence of these activities. First, the absolute number of remedial activities a BT and her/his support team engage in can be used as a measure of the probable remedial impact. For example, it would be assumed ordinarily that reading two good references on a particular discipline problem would influence the BT's behavior more than one good reference. This is a quantitative approach. The second approach assumes that the character of an activity may be associated with its impact on the BT's behavior. For instance, independent study may be more efficient in changing the teacher's behavior than participating in a coaching activity. This is a qualitative approach. This exploratory study used both approaches to determine possible relationships between BT performance scores and professional development plans (PDPs).

Development of Quantitative Measures of Remedial Activities

The first approach for determining the relationship between the number of activities in the professional development plans and scores on pre and post observations entails a quantitative measure of these activities. This approach was developed from data provided by beginning teacher portfolios. The following measures were generated for each portfolio:

1. Observation Measures
    a) Total number of summative observations.
    b) Total number of formative observations.
    c) Combined total of all types of observations: FPMS summative, FPMS formative, and observations using other types of observation instruments.
2. Activities Engaged in
    a) Total number of first assigned activities.
    b) Total number of activities assigned after the first remedial activity (also referred to as multi activities).
3. Total amount of material evidence presented in a portfolio as competency documentation (for example, tests, lesson plans, student papers).
4. Total number of conferences conducted during the year.

Development of Qualitative Measures of Remedial Activities

The second approach involves the value of the remedial activities of the PDP. In order to ascertain an activity's value it was necessary to assign numerical values that could be used to rank order activities on a scale of effectiveness with respect to BT behavior.

A scale was developed for this purpose by guidelines found in research on the effect of various inservice procedures and techniques. Wade (1985) reviewed over 300 journal articles, dissertations, and ERIC documents on the effects of inservice procedures and techniques produced between 1968 and 1983. Ninety-one of these studies were selected for inclusion in a meta-analysis. Only studies which satisfied the following conditions were included: based upon quantitative data, effect-size data were included, subjects were public school teachers, and the data were related to one or more major questions of the study.

Wade's study provided a set of effect sizes associated with various types of procedures and techniques. Table 1 lists practices and their effect sizes, generated from the meta-analysis of 91 research studies. The larger the value, the more effective the practice tended to be in influencing teacher behavior.

TABLE 1

Practices and Their Effect Sizes on Teacher Behavior from a Meta-Analysis of Ninety-One Studies

Practice	Effect Size
Attempts to increase participants' learning	.90
Federal, state, and university initiated programs	.69
Training groups with elementary and secondary teachers	.67
Participants who were selectively chosen to participate	.76
Self-instruction	.92
Independent study	.98
Strong leadership roles by instructors	.66
Instructional techniques:	
Observation	.81
Micro-teaching	.78
Video-audio feedback	.64
Practice	.55

Using the effect sizes and categories reported in Table 1, the types of remedial activities reported in the BT portfolios (Chapter 1, Table 13) were rank ordered. The derived effect sizes and rank values are reported in Table 2.

TABLE 2

Types of Remedial Activities and Their Rank Values

Type of Activity	Effect Size	Rank Value
Professional readings (Independent study)	.98	4
Self-instruction	.92	4
Observing other teachers (Observation)	.81	3
Viewing video tapes (Video-audio feedback)	.64	2
Coaching	**	1
Workshops	**	1
Other	--	2

** Not listed by Wade as one of the more effective activities

Each activity assigned to a beginning teacher was given a rank value from one to four, depending on the effect size of the category. Rank values for coaching, workshops, and "other" were assigned by the research team. Five scores were derived by totaling these values according to evidence found in the portfolio of each BT.

1. Total value of first assigned activity; that is, the combined values of the first activities assigned for each specific domain or competency.
2. Total value of all activities (multi activities) assigned after the first activity is assigned; that is, if more than a single activity was assigned for a specific domain or competency, this value was computed.
3. Average value of the first assigned activities; that is, the arithmetic mean of the values for these activities.
4. Average value of all other activities (multi activities); that is, if one or more activity(ies) was assigned after the first assigned activity for a specific domain or competency, the arithmetic mean of the values for these was computed.
5. Average value; that is, the arithmetic mean of all activity values assigned to a beginning teacher.

Procedure

What is the effect of these various combinations of remedial activities on the behavior of BTs? To explore this question, we applied standard FPMS scoring procedures to produce selected performance scores. Pre and post FPMS summative observations were available in the portfolios for most of the teachers. These observations were scored and seven FPMS subscores were produced: total score, effective score, ineffective score, and scores for each of Domains 2 through 5 inclusive.

Two forms, pre and post, of each subscore, were used for analysis. The first observation subscores were retained as a pretest. A residualized gain score was computed for each of the seven post observation subscores, as a measure of change. This gain score was created by regressing pretest scores onto posttest scores for each BT, then subtracting the predicted from the obtained score to produce a residualized score which controls somewhat for confounding pretest differences (Klitgaard, 1978; Baldwin, Medley and MacDougall, 1983; Micceri, 1986).

Simple Pearson product moment correlations then were computed relating each form of FPMS scores to each of the qualitative and quantitative measures of remedial activities from the portfolios. This resulted in a multiplicity of correlations.

In order to investigate the influence of more than a single remediating activity on teacher behavior, portfolios were divided into three groups.

1. Those having no activities assigned.
2. Those having only a single activity assigned for each competency or domain problem.
3. Those having more than one activity (multi activity) assigned for at least one competency or domain problem.

A oneway ANOVA was run on each of the seven residualized gain scores across these portfolio groups with post hoc multiple comparisons made on those significant at the .05 level (Scheffe). Again, the large number of correlations conducted using this procedure makes this at best an exploratory investigation.

Rank Values of Assigned Remedial Activities

The rank value of activities is one of the most important elements of the PDP. If an activity has a rank value of three or more, it is considered a high-value activity. By interpretation, it can be seen from Table 3 that 50 portfolios, or 37 percent of the 133 cases, include first assigned activities (hereafter called single activities), having an average value of three or more (high-value activities). The activities found in 83 portfolios, or 63 percent, have an average value of less than three (0-2) (lower-value activities). In Table 4, the multi activities show a similar tendency toward low average value. However, there was a large number of portfolios that did not include evidence of activities (n=36 or 27 percent in Table 3; n=91 or 68 percent in Table 4). If the average value is figured on the total assigned activities (n=97 in Table 3; n=42 in Table 4), the number of activities with a value of 3 or greater is 63 or 45 percent (n=50 in Table 3; n=13 in Table 4). This suggests that many assigned activities may have little effect on the teacher's behavior as determined by the high percentage of activities with values less than three.

TABLE 3

Average Value of Single Activities

Average Value	Frequency	Percent
No assigned activities	36	27.1
1.4	10	7.5
1.5	2	1.5
1.6	3	2.2
1.7	1	0.7
1.8	1	0.7
2	17	12.7
2.2	1	0.7
2.2	1	0.7
2.4	1	0.7
2.5	3	2.2
2.6	1	0.7
2.6	4	3.0
2.7	1	0.7
3	20	14.9
3.2	1	0.7
3.3	1	0.7
3.5	1	0.
3.6	3	2.2
4	24	17.9

TABLE 4

Average Value of Multi Activities

Average Value	Frequency	Percent
No assigned activities	91	68.4
1	10	7.5
1.3	1	0.7
1.5	1	0.7
1.7	1	0.7
2	16	11.9
3	11	8.2
4	2	1.5

Effects of Quantitative Remedial Activities

What is the relationship among the quantitative remedial activities and teacher behavior? In other words, is either the effective or ineffective behavior of the BT, as measured by the FPMS, changed by quantitative remediation activities? From Table 5 it can be seen that no dependable connection appears between any of the measures of activities and either effective behavior scores, or total scores. However, ineffective behavior appears to be reduced by each of two types of activities: number of formative observations, and total number of observations (both formative and summative).

More specifically, using minimums of $r = .15$, and $p < .05$, Table 5 suggests a positive and significant relationship between residualized ineffective gain scores and the total number of formative observations ($r = .3862$, $p < .046$); and a positive relationships that approaches significance for the total number of observations of all types ($r = .1695$, $p < .066$).

Why does observation change the BT's behavior? That mere observation of a teacher's classroom behavior would improve that behavior seems unreasonable. How then can this apparent effect of observation be explained? A possible explanation is the likelihood of interaction about one or more specific teacher behaviors. As the BT and one or more members of the support staff review the results of an observation, ineffective behavior may be noted and discussed. This discussion could account for the positive correlations cited above.

Moreover, in considering the effect of remedial activities on effective, ineffective scores or on domain scores it should be remembered that the findings pertain to a sample of 133 portfolios. While this sample was selected from four of the five state geographic regions and from small as well as large districts, still generalizations from this sample should be made with caution.

TABLE 5

Correlations Between PDP Activities And Residualized Gain Scores

Type of Activity	Rsid | Total	Rsid | Effective	Rsid | Ineffective
Number of Formative Observations	0.1536 (n 20) P=0.259	0.0735 (n 20) P=0.379	0.3862 (n 20) P=0.046
Number of Summative Observations	0.0284 (n 78) P=0.403	-0.0252 (n 78) P=0.413	0.1364 (n 78) P=0.117
Total Number of Observations	-0.0132 (n 80) P=0.454	-0.0731 (n 80) P=0.260	0.1695 (n 80) P=0.066
Total Number of Activities in PDP	0.0057 (n 71) P=0.463	0.0193 (n 71) P=0.473	-0.1575 (n 71) P=0.095
Average Value of Activities in PDP	-0.0722 (n 71) P=0.275	-0.0762 (n 71) P=0.264	-0.0486 (n 71) P=0.344
Total Value of Activities in PDP	-0.0114 (n 71) P=0.463	-0.0081 (n 71) P=0.473	-0.1542 (n 71) P=0.100
Total Number of Conferences Conducted	0.0173 (n 98) P=0.433	0.0472 (n 98) P=0.322	-0.1241 (n 98) P=0.112
Total Number of Physically Documented Comps	0.0526 (n 76) P=0.326	0.0730 (n 76) P=0.265	-0.1308 (n 76) P=0.130

Since the absolute number of formative observations conducted for a particular portfolio appears to relate positively to improvement of ineffective behavior ($r=0.3862$, $P<0.046$), the relationship between this variable and the number of activities assigned in a portfolio was investigated. Theoretically, following problem identification by screening observations, formative observations should be conducted to identify appropriate and specific items or problems for remediation. This suggests that a greater number of assigned remediation activities should associate with a greater number of formative observations and vice versa. Contrary to expectation, Table 6 data show that there is no consistent relationship between number of formative observations and number of assigned activities.

TABLE 6

Relationship Between Formative Observations and Activities

	Absolute Number of Activities Assigned			
	None n = 36	One n = 17	Two to Four n = 61	Four n = 19
Average Number of Formative Observations	1.333	1.353	0.885	1.053

TABLE 7

Correlations Between Professional Development Plan Activities and Domain Based Residualized Gain Scores

	Residualized Gain Scores			
Type of Activity	Domain 2	Domain 3	Domain 4	Domain 5
Number of Formative Observations	0.1401 (n 20) P=0.278	0.1452 (n 20) P=0.271	0.2466 (n 20) P=0.147	-0.0535 (n 20) P=0.411
Number of Summative Observations	-0.0096 (n 78) P=0.467	0.0980 (n 78) P=0.197	0.0143 (n 78) P=0.451	-0.1128 (n 78) P=0.163
Total Number of Observations	-0.0534 (n 80) P=0.319	0.0487 (n 80) P=0.334	0.0067 (n 80) P=0.477	-0.1135 (n 80) P=0.158
Total Number of Activities in PDP	0.0523 (n 71) P=0.332	0.0508 (n 71) P=0.337	-0.0443 (n 71) P=0.357	-0.1386 (n 71) P=0.125
Total Value of Activities in PDP	0.0377 (n 71) P=0.377	0.0072 (n 71) P=0.476	-0.0288 (n 71) P=0.406	-0.1322 (n 71) P=0.136
Average Value of Activities in PDP	-0.0504 (n 71) P=0.338	-0.1013 (n 71) P=0.200	-0.0260 (n 71) P=0.415	0.0268 (n 71) P=0.412
Total Number of Conferences Conducted	0.0970 (n 98) P=0.171	0.0026 (n 98) P=0.490	-0.0597 (n 98) P=0.279	0.0454 (n 98) P=0.329
Total Number of Physically Documented Comps	-0.0036 (n 76) P=0.488	0.1343 (n 76) P=0.124	0.0032 (n 76) P=0.489	-0.0268 (n 76) P=0.409

The analysis at the domain level in Table 7 shows no relationship of any type of activity to domain subscores. In Table 7, the number of cases across any type of activity is the same for all domains since the summative instrument was the source of the data. Any case having a residualized gain score for one domain had a residualized gain for all domains.

In order to further investigate the efficacy of having multiple remediation activities for specific problems, the portfolios were divided into three groups.

1. Those cases having no activities assigned in their PDPs.
2. Those cases having a single activity per problem.
3. Those cases having multi activities for at least one domain or competency.

It should be noted that cases in Group 2 could easily have more total activities than cases in Group 3, if Group 3 subjects had only a single identified problem and Group 2 subjects had more than one problem.

Table 8 shows the mean scores per group for each of the seven scores. A one way ANOVA was run across each group for each of the seven residualized scores, with only the results for Domain 4 ($F = 4.094$, $p < .0197$, DF 2.95), attaining significance at the .05 level. Post hoc Scheffe multiple comparison tests found a significant difference between Groups 2 and 3 for Domain 4 scores. This suggests that the assignment of one activity may help to improve teacher behavior but more than one makes little improvement. One might suppose that improvement would occur only if the additional activities for Domain 4.0 rank high in value.

TABLE 8

Number of Activities Per Problem in PDP and Residualized Gain Scores for FPMS Observations

Residualized Gain Scores	Grouped by Number of Activities Per Problem: No Activities Mean (n 27)	One Activity Mean (n 39)	Two or More Activities Mean (n 32)
Total Score	-0.82	1.95	-1.68
Effective Score	-1.01	1.99	-1.57
Ineffective Score	0.22	0.13	-0.35
Domain 2 Score	0.09	0.11	-0.21
Domain 3 Score	-0.65	0.88	-0.52
Domain 4 Score	-0.31	0.79	-0.70*
Domain 5 Score	-0.01	0.20	-0.24

* Significant at .05 level

Relationships between Residualized Gains and Assigned Activities

The preceding analyses dealt mostly with relationships of gross quantitative measures of remedial activities with residualized gain scores of BT behavior. We turn now to a study of the relationship between residualized gains and more specific breakdown of the remedial procedures and techniques. As described earlier, activities are broken into two groups: single activities and multi activities. These groups are somewhat arbitrary. However, they signify whether more than one activity was assigned to remediate a specific problem.

Only 32 portfolios, for which FPMS pre and post scores were available, included more than a single activity in the PDPs; 71 had at least one (Table 9). Since pre and post information was available for approximately 103 cases, 71 percent included one remediation activity and 31 percent two or more for each identified problem, whether a domain or competency was specified as the area of difficulty.

At this more specific level of investigation, some seeming contradictions emerge. Table 9 data show no significant relationships between activities and observations with the exception of the total value of multi activities and the total residualized gain scores ($r = .3089$, $p < .043$). The total

score approaches significance in its negative relationship to the mean value of single activities ($r = -.1933$, $p < .053$) and the effective gain score is negatively related to the mean value of single activities ($r = -.2188$, $p < .033$).

TABLE 9

Relationship Between Performance Residualized Gains and PDP Activities

Type of Activity	Residualized Gain Scores		
	Effective	Ineffective	Total
Total Number of Single Activities	0.1175 (n 71) P=0.165	-0.1716 (n 71) P=0.076	0.0789 (n 71) P=0.256
Total Number of Multi Activities	0.1523 (n 32) P=0.203	0.2445 (n 32) P=0.089	0.2465 (n 32) P=0.087
Total Value of Single Activities	0.0465 (n 71) P=0.350	-0.1545 (n 71) P=0.099	0.0298 (n 71) P=0.403
Total Value of Multi Activities	0.2459 (n 32) P=0.087	0.1599 (n 32) P=0.191	0.3089 (n 32) P=0.043
Mean Value of Single Activities	-0.2188 (n 71) P=0.033	-0.0200 (n 71) P=0.434	-0.1933 (n 71) P=0.053
Mean Value of Multi Activities	0.1897 (n 32) P=0.149	-0.1949 (n 32) P=0.143	0.1207 (n 32) P=0.255

Although these data suggest that the single activities are not effective in changing the behavior of the BT and that the total value of multi activities has a positive influence on the BT's behavior, the relationships must be viewed in light of the data in Table 8 and the data on formative observations. The ineffective scores do not appear to relate much to the number of activities, but the effective and total scores show much greater change, therefore this should not be taken to mean that many activities are necessarily better than one. Notice in Table 8 only one score (Domain 4.0) had a relationship at the significance level of .05. It is clear from the combined data that one activity is definitely better than many, probably because it is specifically related to the problem and a single activity may be easier to monitor.

TABLE 10

Relationship Between Residualized Gains by Domains and PDP Activities

Type of Activity	Residualized Gain Scores			
	Domain 2	Domain 3	Domain 4	Domain 5
Total Number of Single Activities	0.0729 (n 71) P=0.273	0.1115 (n 71) P=0.177	0.0418 (n 71) P=0.365	-0.1018 (n 71) P=0.199
Total Number of Multi Activities	0.1361 (n 32) P=0.229	0.1971 (n 32) P=0.140	0.2872 (n 32) P=0.056	-0.0002\ (n 32) P=0.500
Total Value of Single Activities	0.0620 (n 71) P=0.304	0.0383 (n 71) P=0.376	0.0090 (n 71) P=0.470	-0.1210 (n 71) P=0.157
Total Value of Multi Activities	0.0671 (n 32) P=0.358	0.2397 (n 32) P=0.093	0.5169 (n 32) P=0.001	0.1203 (n 32) P=0.256
Mean Value of Single Activities	0.0352 (n 71) P=0.385	-0.2135 (n 71) P=0.037	-0.1659 (n 71) P=0.083	-0.0988 (n 71) P=0.206
Mean Value of Multi Activities	0.0874 (n 32) P=0.317	0.0884 (n 32) P=0.315	0.3439 (n 32) P=0.027	0.1703 (n 32) P=0.176

Turning now to the domain level, Table 10 shows that Domain 4 is positively and close to significantly associated with the total number of multi activities ($r = .2872$, $p < .056$); significantly associated with the total value of the multi activities ($r = .5165$, $p < .001$); and the mean value of multi activities ($r = .3439$, $p < .027$). It appears from these data that the chances of improving teacher behavior for Domain 4 is enhanced by utilizing multi activities.

It is interesting to note that teacher behavior in Domains 2 and 5 are not related to either the value or the number of activities, while teacher behavior in Domain 3 is negatively correlated to the mean value of single activities.

Summary

The BTP was developed to improve teaching and to provide assistance to the teacher as he/she learns to cope with the problems of the classroom and of adjusting to the school environment. This means that the priority of assistance given to the BT should be geared directly to the difficulties the teacher experiences from time to time. Since these difficulties are generally reflected in the teacher's ineffective rather than effective scores, the ineffective behaviors represented on the FPMS summative instrument will be given special consideration in the teacher's PDP. It is therefore not surprising that the effective behavior of the BT is not improved by the remedial activities of the PDPs.

However, ineffective behavior is changed but the positive changes appear to be induced by interaction, less change occuring where interaction was limited. On the whole, greater improvement occurs when the remedial activities rank higher in effect value and the amount of interaction is high. We should not expect improvement to occur from a mere increase in the number of activities.

Generally speaking, the findings of this study suggest that the classroom behavior of BTs can be improved if the support team conducts multiple observations, identifies ineffective behavior, and induces teachers to engage in high-valued remedial activities. Little improvement can be expected if the activities are of low value even though many are engaged in, nor will improvement likely occur if the activities are not specific to identified ineffective behavior.

In order to improve the quality of the data to be compiled for diagnosis, prescription, documentation, and research, the number of observations should be increased. Observations should be made by certified observers. Beyond the data collected from a study such as the portfolio study reported here, there is need to design studies to collect all observation data, particularly final observations, in order to estimate growth scores of beginning teachers.

In light of the finding that there is no consistent relationship between observations and prescriptions for professional development, a systematic plan for diagnosis, remediation, and documentation should be developed with a program of training in its implementation.

Summary and Discussion

Overall, the effect of the various combinations of activities reported in the portfolio sample can be summarized as follows:

1. An effect scale was constructed, based upon the findings of a meta-analysis study of 91 research studies of the effectiveness of inservice procedures and findings. When the remedial activities, reported in the portfolios, are ranked by this scale, 45 percent of the portfolios have activities of an average value of 3 or above on a scale of zero to four.
2. The effective behavior for this sample of BTs does not appear to be changed by observation, remedial activities, or conferences.
3. The ineffective behavior appears to be improved by the number of formative observations and possibly by the total number of observations (both summative and formative) per teacher.
4. There is apparently no consistent relationship between the number of formative observations conducted and the number of remedial activities assigned per teacher.
5. In this sample, Domain 4.0 is the only domain where the data suggest that the teacher's behavior may be influenced by the number of formative observations, although the relationship only approaches significance. The number of activities engaged per problem seems to positively influence teacher behavior in Domain 4.0. No other domains as identified on the summative instrument appear to be significantly influenced by remedial activity in this study.

More specifically, the scores used in this analysis were all derived from the FPMS summative instrument. Although the total effective and ineffective subscores, and perhaps Domain 3.0 include enough items for stable measurement, it is difficult to extract relationships given the limited number of items in the subscores for Domains 2, 4, and 5 compared to the total number of concepts found in the domains. Thus, even the most modest relationships must be considered worthy of further investigation.

6. The BT performance in Domain 4.0 appears to be improved by the total number of activities in the PDP, the total value of the multi activities in the PDP, and the mean value of the multi activities.

7. Effective scores may be negatively influenced by the mean value of single activities.

CHAPTER 3

Reporting Information to School Districts and Colleges on the Performance of Beginning Teachers

The subject of the first two chapters of this report is the status and impact of the BT portfolios on teacher behavior. The first one deals with the content of the portfolios. The findings of this study indicate not only the need for a more systematic plan of record keeping but also for districts to be more diligent in collecting and using the data called for by State Board Rule. The second chapter assesses the impact of the activities reported in the portfolios on the behavior of BTs. The information presented in this chapter depicts the use of a large number of remedial activities whose influence on teacher behavior is weak. It also shows a lack of coordination between formative evaluation and assignment of remedial activities. While the effect of these activities on BT behavior is much less than satisfactory, the results of the assessment seem to suggest that improvement is likely to occur if the support team makes multi observations, diagnoses teacher behavior, and engages teachers in at least one relevant remedial activity of high effect value.

A Proposed Feedback Plan for Districts

The present chapter sets forth a plan for providing feedback to districts, colleges, and the State Department of Education on performance of BTs. In order for colleges to assess the effectiveness of their preservice programs and for districts to assess their BT training programs, it is necessary that they have information derived from valid and reliable observations and evaluations of the teachers they train. Teacher evaluation, however, is comparative. There presently exists no absolute values, set by validated measures of performance, that separate effective from ineffective teachers. Instead, we are dependent upon normed scores on which standards can be set.

In order to develop norms for performance measures and to maintain current norm scores, accurate evaluations of beginning teachers across the State must be analyzed and feedback from these analyses submitted to districts, colleges, and the State Department of Education on a regular schedule. Through this type of procedure the comparative performance of beginning teachers will become a matter of record and can be used to systematically improve teacher training programs.

The State Department of Education has taken initial steps to develop a system of evaluation, procedures for analyzing evaluations, and a data processing and analyzing agency for the foregoing purposes. The teacher evaluation system that has been validated and for which beginning teacher norm scores have been created is the Florida Performance Measurement System. The DOE has collected a substantially large number of screening and summative FPMS evaluations of beginning teachers and sent them to the Teacher Evaluation and Assessment Center. At the TEAC, the data have been entered into automated files for analysis and work has been initiated to develop formats for reporting evaluation results to districts, colleges and the DOE. This chapter reports possible formats and examples of information that could be reported on a regular basis providing provision is made in State Board Rule for the collection of screening and summative evaluations of beginning teachers.

Several kinds of data are of value to districts, colleges, and the DOE. The kinds of data referred to in this chapter are measures of classroom performance that, according to the literature on teacher effectiveness, are essential in determining teacher effectiveness or ineffectiveness in promoting student growth. In working with teachers to improve student performance -- behavior and achievement -- this is the most relevant kind of information.

The proposed system for providing feedback to districts, colleges, and the DOE is designed to target behaviors that teachers do well and those they do poorly. It may be used to provide such comparative information by district, school, or individual.

Through the use of a computerized scoring and reporting system, developed for the FPMS and normed on beginning teachers, it is now possible to provide the foregoing information. This system is designed to provide for a compilation of all FPMS scores for all BTs in all Florida school districts that use the FPMS in their beginning teacher programs.

Table 1 is a prototype of a format for reporting average evaluation scores of a district's BTs based upon state norm scores. The data in this table are modeled from actual results, and represent all schools providing information.

The table lists the number of teachers included in the analysis, and is divided into four segments. The first segment reports a total district percentile score, a total effective score, and a total ineffective score. The second segment reports domain scores. Segment three reports item

scores for effective indicators and segment four reports scores for ineffective indicators. All of this information is initially derived from observations of teachers in classrooms using the FPMS summative instrument.

TABLE 1

Prototype Format for Giving Feedback to Districts on Beginning Teacher Performance

District: Okadata — Number of Cases: 152

	FPMS Scores	Mean Percentile Rank	Graph of Mean Percentile Rank 25 50 75
	----- TOTAL SCORES -----		
	Total FPMS Score	52	**********
	Total Effective Score	53	**********
	Total Ineffective Score	53	**********
	----- DOMAIN SCORES -----		
	Domain 2 - Conduct	60	**********
	Domain 3 - Instruction	35	**********
	Domain 4 - Subject Matter	59	**********
	Domain 5 - Communication	59	**********
	----- EFFECTIVE INDICATOR SCORES -----		
1.	Begins Instruction Promptly	93	**********
2.	Handles Material Orderly	46	**********
3.	Orients Stud./Maint. Focus	68	**********
4.	Beginning/Ending Review	70	**********
5a.	Single Factual Questions	58	**********
5b.	Requires Analysis/Reasons	67	**********
6.	Rec. Resp./Correct Feedback	73	**********
7.	Specific Academic Praise	29	**********
8.	Provide for Practice	69	**********
9.	Hw/Stwork Assign/Gives Dir	72	**********
10.	Circul. & Assists Students	67	**********
11.	Concept Treatment	72	**********
12.	Discuss Cause-Effect	77	**********
13.	State/Apply Academic Rule	72	**********
14.	Crit/Evid for Value Judge	84	**********
15.	Emphasize Important Points	66	**********
16.	Express Enthus./Challenge	69	**********
19.	Body Behav. Shows Interest	70	**********
20.	Stops Misconduct	70	**********
21.	Maintains Instruct Momentum	71	**********

Table I (continued)

FPMS Scores		Percentile Rank	Mean Percentile Rank 25 50 75
	----- INEFFECTIVE INDICATOR SCORES -----		
1.	Delays	93	**********
2.	Not Organized/Materials	72	**********
3.	Talk Unrelated to Topic	83	**********
5a.	Mult Quest/Unison Response	76	**********
5b.	Non-academic questions	88	**********
6.	Ignoring, Sarcasm, Disgust	95	**********
7.	General, Nonspecific Praise	73	**********
8.	Extends Discourse	98	**********
9.	No Homework/Feedback	93	**********
10.	Circulates Inadequately	93	**********
11.	Gives Def/Examples Only	94	**********
12.	Cause or Effect Only	97	**********
13.	Does Not St/Apply Acad Rule	98	**********
14.	Value Judge With No Crit	97	**********
17.	Vague, Scrambled Discourse	96	**********
18.	Grating/Inaudible Talk	92	**********
19.	Deadpan/Lethargic	96	**********
20.	Delays Desist/Punitive	79	**********
21.	Loses Momentum	85	**********

Table 1 shows that the total FPMS score for the teachers in this district is at the state normed 52nd percentile. The total effective score and the total ineffective score (53,53) indicate that the beginning teachers in this district are about average for the state. For all scores on FPMS, the higher the score, the more effective the performance.

The second group of scores reported in Table 1 are for each of the four domains measured on the FPMS summative instrument. In general, the teachers in this district scored above average in all of the domains. The lowest score, however, is Domain 3, Organization and Development of Instruction. Efforts to improve teacher performance in the organization and development of instruction may be considered the district's priority area of training since it is the lowest of the four domain scores. To make these efforts most productive, it is necessary to identify the specific behaviors of Domain 3 that most need improvement in this school district. These can be seen by singling out the lowest scores of the effective indicators as follows:

1.	Specific academic praise	29	pct. rank
2.	Handles materials orderly	46	pct. rank
3.	Asks single factual questions	58	pct. rank
4.	Require analysis/reasons	67	pct. rank
5.	Circulates and assists students	67	pct. rank
6.	Orients students/maintains focus	68	pct. rank
7.	Provides practice	69	pct. rank

By referring to Table 1 again, it can be seen that this district's ineffective percentile scores are very high, which means few ineffective behaviors were recorded. This indicates that there is no pressing need for general remediation of ineffective teacher behavior in this district, however, improvement could be made in items 2) Does not organize or handle materials systematically; 5a) Poses multiple questions/unison response; and 7) general/nonspecific praise.

It should be emphasized that these statistics represent averaged percentile ranks for all teachers of this district. They do not pertain to individual teachers per se. In order to make specific diagnoses of individual teachers, one must analyze observations of individual performance using the summative instrument as a means of detecting general areas of instruction that might be improved, and then make follow-up formative observations for purposes of specific diagnoses and planned remediation.

A Proposed Feedback Plan for Colleges

Like the districts and the DOE, the colleges and departments of education of the State's higher institutions of learning need to know how well BTs perform in the classroom. The DOE is responsible for deciding whether or not to accredit particular college programs of teacher education, and the colleges, on the other hand, need to know how well their BT graduates perform on the job in order to improve their programs of teacher preparations and to satisfy the state requirements for accreditation. One of the most dependable ways of determining the strengths and weaknesses of a college's products is to measure the initial performance of its BT's classroom behavior.

Each BT's performance, according to State Board Rule, must be evaluated within the first forty-five working days of employment, and the evaluation must be based upon a state approved observation instrument. These data properly collected, analyzed, and reported make it possible for the DOE to supply college and departments of education with the means of assessing their products.

Through the use of a system of scoring and reporting, developed for the FPMS and normed on beginning teachers, the DOE can now supply the college or departments of education with the appropriate information on their BTs. Table 2 displays a format for providing feedback on BTs to the State's teacher education institutions.

Table 2 format parallels the one proposed for district use. It provides FPMS scores, mean percentile ranks, and graphic representation of the mean percentile ranks. Like the district format, it also provides total scores, domain scores, effective indicator scores, and ineffective indicator scores.

From Table 2, it can be seen that the average total FPMS score for the BT graduates of this college, who entered teaching that year, is at the 24^{th} percentile of the state normed distribution. The total effective score (25) indicates that the BTs of this college are much below average for the state. According to the ineffective total score (47), the same teachers are slightly less than the mean percentile.

TABLE 2

Prototype Format for Giving Feedback to Colleges on Beginning Teacher Performance

Institution: Calftoy

Number of Cases: 140

	FPMS Scores	Mean Percentile Rank	Mean Percentile Rank 25 50 75
	----- TOTAL SCORES -----		
	Total FPMS Score	24	*****
	Total Effective Score	25	*****
	Total Ineffective Score	47	**********
	----- DOMAIN SCORES -----		
	Domain 2 - Conduct	23	*****
	Domain 3 - Instruction	25	******
	Domain 4 - Subject Matter	44	*********
	Domain 5 - Communication	49	*********
	----- EFFECTIVE INDICATOR SCORES -----		
1.	Begins Instruction Promptly	87	********************
2.	Handles Material Orderly	45	**********
3.	Orients Stud./Maint. Focus	61	**************
4.	Beginning/Ending Review	56	************
5a.	Single Factual Questions	51	**********
5b.	Requires Analysis/Reasons	54	**********
6.	Rec. Resp./Correct Feedback	61	**************
7.	Specific Academic Praise	46	**********
8.	Provide for Practice	38	********
9.	Hw/Stwork Assign/Gives Dir	62	**************
10.	Circul. & Assists Students	40	********
11.	Concept Treatment	55	**********
12.	Discuss Cause-Effect	77	******************
13.	State/Apply Academic Rule	67	****************
14.	Crit/Evid for Value Judge	80	*******************
15.	Emphasize Important Points	73	*****************
16.	Express Enthus./Challenge	49	**********
19.	Body Behav. Shows Interest	63	**************
20.	Stops Misconduct	50	**********
21.	Maintains Instruct Momentum	58	**********

Table 2 (continued)

FPMS Scores		Percentile Rank	Mean Percentile Rank 25 50 75
	----- INEFFECTIVE INDICATOR SCORES -----		
1.	Delays	86	**********
2.	Not Organized/Materials	100	**********
3.	Talk Unrelated to Topic	64	**********
5a.	Mult Quest/Unison Response	53	**********
5b.	Non-academic questions	76	**********
6.	Ignoring, Sarcasm, Disgust	100	**********
7.	General, Nonspecific Praise	93	**********
8.	Extends Discourse	100	**********
9.	No Homework/Feedback	87	**********
10.	Circulates Inadequately	73	**********
11.	Gives Def/Examples Only	87	**********
12.	Cause or Effect Only	100	**********
13.	Does Not St/Apply Acad Rule	100	**********
14.	Value Judge With No Crit	100	**********
17.	Vague, Scrambled Discourse	100	**********
18.	Grating/Inaudible Talk	100	**********
19.	Deadpan/Lethargic	100	**********
20.	Delays Desist/Punitive	65	**********
21.	Loses Momentum	100	**********

When the performance of the graduates of this college, for the year under consideration, is viewed by domains, the second group of scores indicate that these BTs are dramatically below average (23 percentile) in the management of student conduct and only slightly better (25 percentile) on the organization and development of instruction. They approach the mean percentile rank for Domain 4 (44 percentile) and for Domain 5 (49 percentile).

It is also clear from a review of the effective indicators, Table 2, that as a group these BTs are below the state average in the performance of certain indicators -- 2) handling materials orderly, 7) specific academic praise, 8) providing practice, 10) circulating and assisting, and 16) enthusiasm and challenge. On these five indicators, they rank close to or below the mean 50th percentile among the BTs of the State's colleges. The BTs of this college were scored high on ineffective indicators except for three which were only slightly above the 50 mean percentile; namely, 3) table unrelated to topic, 5a) multi questions and unison response, and 20) delays desist and punitiveness. All in all it seems evident that some change in the program of this college may be beneficial in raising the average score of its graduates.

Comparative Format for Colleges

There is a question of policy that should be addressed by DOE. It concerns the distribution of information about the comparative performance of the BTs of the various Florida colleges and out-of-state institutions. Should the data on the performance of BTs of each college be made available by DOE to each college only or should the data be available to each and every college alike, thereby making it possible for each college to compare its products with those of others? In addressing this question, it is important to focus the debate on the purpose of these data, which is to aid in the improvement of education. How these data are circulated and used must be to this end. Use of these data for political purposes would be inappropriate and could jeopardize the system.

Timing of the release of this information is an important factor. Time must be allowed for each district to guarantee that all screening observations have been made by trained, certified, and updated observers using valid and reliable observation instruments as approved by the State. In addition, the number of observations must be sufficient to provide a representative sample of the teacher's performance. Unless these conditions are met, comparative information will breed controversy that could threaten the whole system of evaluation.

Assuming that the question of policy has been settled and comparative data are to be made available to the colleges, Table 3 is offered as an initial format subject to modification as experience warrants.

Table 3

Relative Average Percentile Rankings by Colleges

Cases	Total Score	Total Effec. Score	Total Ineffec. Score	Domain 2 Score	Domain 3 Score	Domain 4 Score	Domain 5 Score
7	24	24	47	26	25	44	49
22	37	39	41	48	45	51	46
161	50	50	55	56	52	54	57
106	51	51	63	58	53	54	56
56	52	52	58	61	55	57	54
85	55	54	63	58	53	54	56
156	49	49	59	58	49	54	52
14	54	53	63	46	56	62	65
150	53	53	59	55	57	56	59
26	64	65	63	66	63	64	65
14	62	60	62	74	61	51	68

Table 3 shows the total FPMS score, total effective and total ineffective scores, and the total screening scores for each of Domains 2 to 5 inclusive. The percentile scores enable a college to determine its relative standing among the Florida colleges with respect to the performance of its graduates who enter teaching a given year. The accumulated rankings across time tells the story of a given college's improvement gains or losses.

Summary

The development of feedback procedures from the DOE to the colleges and districts is an essential feature of a properly functioning BT program. In the foregoing pages we have suggested an initial set of feedback formats and illustrated how they can be used by school districts, colleges, and the Department of Education for the evaluation and improvement of the Florida BT program.

DEVELOPMENT WORK NEEDED FOR BEGINNING TEACHER PROGRAM

Our studies of the BTP, and apparent restrictions on resources and time, have led us to consider what steps to take to maximize improvement in the program. Accordingly, we propose that the following projects be undertaken as early as possible.

Proposed Centralized Data Analysis

For colleges and districts to assess the effectiveness of their respective preservice and inservice teacher training programs and for the state to monitor the quality of teachers education in Florida, it is necessary that they have information derived from valid and reliable performance measurement and evaluation of beginning teachers. Such information should include the strengths and weaknesses demonstrated by beginning teachers in the classroom.

The present state of the art of measurement and evaluation does not provide reliable criterion measures of teacher effectiveness. Instead of such criterion measures, we must depend upon normed distributions that provide ranges of scores on which standards can be set.

To norm performance measures and keep norm scores current, accurate evaluations of beginning teachers across districts must be analyzed on an on-going basis and feedback from these analyses must be submitted to districts, colleges of education, and the Department of Education. It is through this type of procedure that the comparative performance of beginning teachers will become a matter of record that can be used to systematically improve teacher training programs whether they be pre or inservice programs.

The Department of Education has taken initial steps to develop a system of evaluation, procedures for analyzing evaluations, and the creation of a data processing and analyzing agency for the purposes outlined above. The Teacher Evaluation and Assessment Center at the University of South Florida has entered beginning teacher observation data into automated files for which formats for analysis and reporting FPMS scores are being developed and tested.

These data serve as an objective and accurate basis upon which to work with individual teachers in the improvement of instruction, target inservice training needs for the general improvement of instruction within districts and give the state common measures of the quality of performance of beginning teachers.

We therefore, recommend:

1) That updating of state norm scores based upon beginning teacher classroom performance observations be continued and the procedures be further developed, so that the DOE and school districts can monitor individual as well as grouped performance scores of beginning teachers as a means of measuring instruction on a longitudinal basis,
2) That the proposed formats be refined and used for reporting beginning teacher data to teachers, districts, and the DOE.

Proposed Diagnosis, Remediation, and Documentation Plan

To improve the effectiveness of the program, a plan should be developed for targeting competencies that are being performed the least effectively. This will involve a study of procedures for establishing more direct relationships between observation data on BT competencies and district inservice activities. We therefore propose that the following steps be taken:

1) Develop a plan for administering formative observation instruments and other measurement devices that will aid support teams in the job of matching performance data with the demonstration of competencies,
2) Develop a set of peer teacher films on procedures of diagnosis and prescription,
3) Learning packages and other high-effect-valued materials keyed to the competencies should be identifiedand developed for use in the development of the PDP),
4) Develop a program of training (if this is not already planned) for peer teachers and other support team members in systematic diagnosis, prescription, and documentation of beginning teacher performance,

5) Develop a state-wide percentile-range of total performance scores and select criteria that define the presence or absence of each competency. In determining adequate performance for certification, a minimum total performance score should be attained, and each competency should be documented as having been performed using specified criteria.

CHAPTER 4

Predictive Validity Study Designs and Findings

The purposes of this component of the 1985-1986 research on teacher evaluation were to, 1) reanalyze data collected in the 1984-1985 FPMS Predictive Validity Studies using the procedures suggested by the review team, 2) draft predictive validity study designs for examining teacher effectiveness at the specific level of teachers in classrooms and at the more general level of school effectiveness within districts, and 3) to conduct an exploratory study using the general level design.

The Predictive Validity Studies, conducted in 1984-1985, that examined the relationships between FPMS scores and measures of student task engagement and achievement, were reviewed by a team of outside experts. This team was composed of Patricia M. Kay of Baruch College, City University of New York, Harold Mitzel of Pennsylvania State University, and Ronald D. Adams of Western Kentucky University. Results of this review prompted reanalysis of these data using residualized gain scores (RGS) as the dependent variable. That reanalysis was conducted and is found in Section I of this chapter.

This section includes comparative analyses of four studies (elementary social studies, secondary American History, and two studies of elementary mathematics) using the RGS method of computing achievement gains and the method of using average stratum posttest (ASP) scores. Findings across all studies show a positive and significant relationship between FPMS scores and student task engagement, student task engagement and student achievement, and FPMS scores and student achievement. In separate analyses of these studies, the only non-significant relationship found was between FPMS scores and student achievement in the secondary study.

The review team also prompted the development of guidelines for conducting predictive validity studies for measuring teachers' classroom performance. These guidelines are intended for use by researchers in developing proposals for conducting validity studies on teacher evaluation systems and for use by persons reviewing these proposals for purposes of funding. The guidelines are found in Section II of this chapter.

Section III of this chapter contains a brief discussion on design considerations for conducting macro predictive validity studies using school-level teacher evaluation scores as the independent variable and school-level student achievement scores as the dependent variable. This discussion is based upon data collected on teacher performance from the Master Teacher Program and general measures of student achievement in one Florida school district.

Section I

A Meta Analysis of Four FPMS Predictive Validity Studies

The following is a combined studies reanalysis, as suggested by Kay, Evertson and Mitzel (1985), of four predictive validity studies conducted on the FPMS. This approach allows the combining of outcomes thereby increasing the power of test statistics. It is particularly useful when small numbers of cases having small but consistent effects are under study. This frequently occurs when one uses the class rather than the individual as the unit of analysis in a study.

Four studies are included in this analysis conducted in the areas of mathematics, social studies and American Literature. Three of these studies were conducted at the elementary level, and one at the secondary level.

The major studies are identified as A, B, and C. Study A contains two substudies, one in mathematics and the other in social studies. All studies were conducted in regular classrooms and included the following grade levels: 2, 3, 5, 6, 8, and 11. Study length varied from two weeks to one academic year. Specific results for each study are found in a series of unpublished papers (Florida Performance Measurement System: Predictive Validity Studies A-D).

Student achievement was used as a dependent variable in all four studies. Student task engagement, the number of students on task at a given time, is generally considered a mediating variable between teacher performance and student outcomes. Student task engagement was investigated for its direct relationship to both FPMS scores and student achievement measures,

The relationship between FPMS scores and both student achievement and task engagement were positive across all four studies. These relationships were significant in all cases except between FPMS scores and achievement at the secondary level. The relationship between student task engagement and student achievement was positive and significant in all three studies. These results help to confirm the findings of the research on teacher effectiveness that shows a relationship between the items on the FPMS instrument and measures of student outcome (Appendix B).

Primary Research Questions

The studies conducted for purposes of testing the validity of the FPMS summative instrument are primarily concerned with a single research question:

1. Do FPMS scores relate positively to measures of student outcome?
    a) Such outcomes include student learning activities measured by student task engagement and student learning measured by a variety of achievement tests.

To answer the above question, several studies were conducted during the 1984-86 school years. The relationship between FPMS scores and student outcomes were estimated in four studies covering three subject areas (math, social studies and American literature) at several grade levels (2,3,5,6,8,11) in different geographical and demographic areas of the state of Florida (two urban districts, one medium sized district,and one rural district) and for different lengths of time (from two to thirty-two weeks).

Designs and Variables

A variety of designs and controls are implemented in the various studies. Table 1 shows characteristics of the various studies. In all studies, teachers were instructed to teach a "normal" lesson when observed. Only in Study A did any teachers receive training on FPMS related behaviors. Dependent variables (post tests) for Study A were closely tied to teacher objectives in a two week lesson plan. Dependent variables in Study B were based on district-wide objectives. Dependent variables in Study C were only indirectly related to teacher objectives. For all studies, at least two observations by two trained and certified observers were conducted on two occasions. FPMS scores were based on the average of these two observations. The use of two observers is based on data from the FPMS reliability study which indicated that gains in reliability/objectivity resulted from the use of at least two different observation sources (Micceri, 1983). Since each observer in each study observed multiple teachers, no interaction effects between observers and regional peculiarities (socio-economic status, race, etc.) should be expected in the data.

Different studies had the following specific characteristics.

1. Study A - conducted at the second, third, and fifth grade levels in 51 classes in two central Florida school districts. Mathematics or social studies lessons were taught by teacher interns over a two week period. Locally developed pre and posttests, as well as two covariates--the Comprehensive Test of Basic Skills (CTBS) and general reading and mathematics subscales--were available. A measure of student task engagement defined as the average percentage of students on task during observations was obtained as

an estimate of student learning activities. The Teacher Education Internship Project final report (1984) describes the locally developed tests:

> Six unit tests, one for each of the instructional units, were developed and pilot tested. Each test had parallel forms. One form was used as the pre-test, the other as the post-test. Tests were developed based on the content of the units and the learner objectives identified.
>
> The Grade 2 Social Studies test contained 18 items, the Grade 3 and Grade 5 tests contained 22 items. Each of these tests contained true/false and multiple choice items. The mathematics tests contained multiple choice and completion-type questions in which the student was asked to supply the correct answer. The tests at all three grade levels were 20-item tests. The maximum score for the Grade 2 and Grade 3 tests was 20, while the maximum score for the Grade 5 test was 27. The tests were reviewed to assess their content validity, and pilot tested. Data obtained from the pilot testing were used to make revisions where necessary.

Note - means and standard deviations across grade levels were almost identical on both pre and post-tests.

2. Study B - a one semester course taught in the 8th and 11th grades in 40 American History classes in a southern Florida school district. Two observations were conducted on each teacher, and the California Achievement Test (CAT), general verbal test, was used as a covariate. A measure of student task engagement defined as the average percentage of students on task during observations was obtained as an estimate of student learning activities.

   The posttest in this study was a district normed, 40 item American History test based on the objectives of American History lessons taught in the 8th and 11th grades. This test was in its third year of development, and two thirds of the test questions were operational at the time of the study. Higher order cognitive questions were not included on this test since they were being piloted and were scheduled for inclusion the following year. Separate forms and norms exist for both 8th and 11th grade.

3. Study C -- A year long study conducted in 11 schools, 35 classes with 27 teachers. The pretest consisted of Spring 1984 administration of the 30 item Comprehensive Assessment Program Mathematics Subtest, and the posttest was the Spring 1985 administration of the same test. This is a very general test bearing no direct relationship to teacher objectives. Two covariates--CAP reading and language tests--were also available. Each teacher was observed by at least two observers. Most teachers were

observed twice in the Fall 1984 and twice in the Spring 1985. Some teachers taught more than a single class. For purposes of analysis, teacher means were used as the unit of analysis to reduce dependency. This, of course, also reduced the degrees of freedom associated with tests of significance.

Table 1 shows the characteristics of studies included in this investigation of the predictive validity of the FPMS summative observation instrument, using meta analytic procedures and techniques.

TABLE 1

Overview of FPMS Predictive Validity Studies

	Study Number		
	A_1 & A_2	B	C
Number of Classes	18 31	40	34
Number of Teachers	18 31	40	26
Grade Levels	2,3,5	8,11	6
Subject Areas	Math(n=18) Soc Sci(n=31)	American History	Mathematics
Post Test	Content Specific	Content Specific	General Achievement
Pretest	Parallel Form	-----	Parallel Form
Covariates	CTBS Math and Reading	CAT general Language	CAP Lang., Reading
Estimate of Socio Economic Status	Free Lunch Data	-----	Free Lunch Data
Length of Course	two weeks	one semester	one year
Other Frame Factors	school size	school size	school size
Student Activities	task engagement	task engagement	-----

Analyses

In this analysis, simple Pearson product correlation coefficients were computed between the independent variable (FPMS scores) and two adjusted forms of the dependent variable (achievement) explained below. Student task engagement was also assessed in relation to both FPMS scores and student achievement scores. For all analyses and variables, teacher means were used as the unit of analysis.

Studies designed to compare teacher performance must conduct such comparisons using similar students for each teacher. Since all of these studies work with intact classes in which neither students nor teachers are randomly assigned, two techniques were used to control the variability between classes and to create adjusted dependent scores upon which correlation coefficients with FPMS scores could be computed:

1. Class mean raw posttest scores for the average stratum of students (ASP) -
   a) Student scores on the pretest were used to divide the student sample (n=734) into four quarters. Scores for students falling in the middle two quarters (n=362) were retained for analysis. The class arithmetic mean posttest score (ASP) for these "average" students was used as the dependent variable. A minimum of three observations per class was required for inclusion in analysis. Use of the posttest scores of students attaining average pretest scores assures that teachers are compared on similar groups. For more detail on the average stratum concept see validity report 1 (Micceri, 1985).
2. Residualized gain scores (RGS) -
   a) Following recommendations by Kay et. al. (1985). and using a model similar to that of Klitgaard (1978), residualized gain scores (RGS) were computed for each student in the total sample (Study A, n=1348; Study B, n=1112; Study C, n=734). The model used to develop residualized gains was a linear regression of several predictor variables including a pretest or covariate onto posttest mathematics scores to produce individual predicted scores. Posttest scores for each student were then subtracted from the predicted scores to produce residualized gain scores (RGS) and the arithmetic mean for each class was computed to obtain dependent variables with the class mean as unit of analysis.

Each of the two forms of the dependent variable (RGS and ASP) contained different sources of error (Cook and Campbell, 1979) and therefore, use of one dependent measure may be considered a cross-validation of the correlation coefficients between independent and dependent variables. Correlations were computed between FPMS scores and both RGS and ASP, between FPMS and task engagement, and between task engagement and RGS and ASP. In addition, a multiple regression model using both FPMS scores and task engagement to predict RGS was applied. Available predictor variables differed for each study.

1. Study A - individual pretest score, class mean pretest score, CTBS general language subtest, CTBS general mathematics subtest, school size and a school based estimate of socio-economic status (multiple r = .49, n = 1348).
2. Study B - individual covariate score, class mean pretest score and school size (multiple r = .64, n = 1112).
3. Study C - individual pretest score, class mean pre-test score, CAP reading and language, school size and a school based estimate of socio-economic status (multiple r = .75, n = 734).

The quantitative approach to multiple studies provided by meta analytic techniques allows one to combine several outcomes, thereby increasing the power of test statistics. Maximizing power was particularly useful for studies having small numbers of cases or when a program had a small but consistent effect.

Rosenthal (1978) suggests the use of a combined Z based upon the one-tailed p values of specific relationships to test the overall significance of results. The combined Z is computed using the following formula:

$$Z = z / N$$

where z = the z associated with a specific one-tailed probability, and

N = the number of studies

Combined Z scores across one-tailed probabilities associated with correlations were computed across the following groupings of studies.

1. All studies combined (n=4),
2. All elementary studies (n=3),

3. All mathematics studies (n=2),
4. All social studies/language arts type studies (n=2).

Outliers

Three outliers were excluded from these analyses, two in Study B and one in Study C. Outliers were identified using the following logic.

1. Any standardized score lying 3.00 or more standard deviations above or below the mean for any of the key variables (engagement, FPMS, achievement) was investigated.
2. If a substantial distance (one or more standard deviations) separated this score from the next highest or lowest score, this individual was defined as an outlier, and excluded from analysis since such deviations exert disproportionate influence on variance estimates and tests of significance. Two individuals fell into this category.
3. Bivariate plots of key variables were investigated to identify bivariate outliers. One individual fell into this category.

Elimination of these outliers created some change in correlation coefficients.

1. Study B -
   a) Increase in the relationship between task engagement and achievement,
   b) Increase in the relationship between FPMS scores and achievement, and
   c) Decrease in the relationship between FPMS scores and task engagement.
2. Study C -
   a) Decrease in the relationship between FPMS scores and achievement.

Details on specific cases are included in Appendix B.

Relationship Between FPMS Scores and Achievement

Appendix B contains the specific sources for the results contained in Table 2. As the table shows, both versions of the dependent variable produced nearly identical results over a group of four studies. This suggested that they both estimated the same underlying characteristic, which may be defined as the relationship between measures of teacher behavior and measures of achievement for the "average student."

TABLE 2

Combined Z Scores Between FPMS Scores and Achievement

Category	N of Studies	Combined Z Scores for Relationships Between FPMS Scores and Two Forms of Student Achievement: Residualized Gain Scores Z Score	Combined Z Scores for Relationships Between FPMS Scores and Two Forms of Student Achievement: Average Stratum Posttest Z Score
All Studies	4	2.50 ***	2.49 ***
Mathematics	2	2.19 **	2.06 **
Social Std.	2	1.34 *	1.47 *
Elementary	3	2.44 ***	2.56 ***
Post-Elementary	1	0.77	0.56

*** Significant .01
** Significant .05
* Significant .10

Both measures attained a relationship significant at the .01 level between FPMS scores and achievement across all studies (n=4) and for elementary studies (n=3). Both attained relationships significant at the .05 level for mathematics classes (n=2), at the .10 level for social studies/history type classes (n=2), and a non-significant relationship at the post-elementary level (n=1).

Relationship Between FPMS Scores and Task Engagement

Each individual relationship between FPMS scores and task engagement was significant at or about the .01 level. Therefore, only simple correlations are reported in Table 3 on the following page.

TABLE 3

Correlations Between FPMS Scores and Task Engagement

Category	N of Cases	Task Engagement: Pearson r	Task Engagement: P
Study A			
Mathematics	21	.42	.029
Social Std.	31	.44	.006
Study B			
Am. History	38	.65	.0001

Relationship Between Task Engagement and Achievement

Table 4 shows the combined Z scores and resultant probabilities among three studies between measures of student task engagement and two estimates of achievement (residualized gain scores and average stratum posttest scores). Both measures attained a relationship significant at the .01 level across all studies (n=3), and for social/history studies. Both attained a relationship significant at the .05 level for elementary studies (n=2), with the residualized gain scores attaining the .01 level. Disagreement occured on the single mathematics study, with residualized gain scores attaining a .01 level of significance and average stratum posttest scores having a non-significant Z. This, of course, was based on only one study, so one might expect a mid-point to be more accurate than either individual estimate. For both elementary level studies (n=2) and post-elementary level studies (n=1), the average stratum posttest scores attained a slightly lower level of significance than did the residualized gain scores (precisely: .0005 to .0445 for elementary, and .044 to .052 for post-elementary).

TABLE 4

Combined Z Scores Between Task Engagement and Achievement

Category	N of Studies	Combined Z Scores for Relationships Between Task Engagement Scores and Two forms of Student Achievement	
		Residualized Gain Scores Z Scores	Average Stratum Posttest Z Scores
All Studies	3	3.70 ***	2.32 ***
Mathematics	1	2.41 ***	0.27
Social Std.	2	2.83 ***	2.64 ***
Elementary	2	3.32 ***	1.69 **
Post-Elementary	1	1.71 **	1.62 *

*** Significant .01
** Significant .05
* Significant .10

Shared Variance Between FPMS Scores and Task Engagement

The following discussion considers the regression of both FPMS scores and task engagement measures upon student achievement measures in Studies A and B. These regressions are conducted upon both RGS and AST.

Table 5 shows partial correlations between FPMS scores and achievement measures with the variance accounted for by task engagement partialed from both variables. Correlations are shown for task engagement with the FPMS scores partialed out. Both multiple r's and simple r's are included for comparison.

As the table shows, the relationships between FPMS scores and achievement were positive for elementary, and negative for post elementary with the influence of task engagement partialed out. For task engagement, with the variance accounted for by FPMS scores eliminated, the relationship with achievement was positive and above .24 in all cases. These results suggest that much of the variance accounted for by either FPMS or task engagement in the achievement measure was "shared variance".

TABLE 5

Partial Correlations of Dependent and Independent Variables

			Correlations of FPMS Scores and Student Task Engagement with Achievement			
			Residualized Gain Scores		Average Stratum Raw Posttest	
Study	N	Score	Partial r	Simple r	Partial r	Simple r
Study A Mathematics	16	Multiple r		.66		.31
		Engagement	.247	.54	.269	.16
		FPMS Score	.476	.57	- .027	.31
Study A Social Science	30	Multiple r		.43		.40
		Engagement	.375	.43	.301	.39
		FPMS Score	.021	.22	.125	.28
Study B American History	38	Multiple r		.29		.29
		Engagement	.262	.28	.272	.27
		FPMS Score	-.073	.13	-.107	.10

Findings and Discussion

Three of the four studies related scores on the FPMS summative instrument to measures of student learning activities (task engagement) and all four studies related FPMS scores to student achievement. The following conclusions may be drawn from the data.

1. Across all studies (n=4), a significant relationship ($r = .25, p < .006$) appears between FPMS scores and measures of student achievement.

2. Across all studies (n=3), a significant relationship ($r = .50$, $p < .001$) appears between FPMS scores and measures of student task engagement.
3. Across all studies (n=3), a significant relationship ($r = .35$, $p < .001$) appears between student task engagement and measures of student achievement.
4. In elementary classes (n=3), a significant relationship ($r = .31$, $p < .006$) appears between FPMS scores and student achievement.
5. In post-elementary classes (n=1), a non-significant relationship ($r = .11$, $p < .255$) appears between FPMS scores and student achievement.
6. In mathematics classes (n=2), a significant relationship appears ($r = .33$, $p < .018$) between FPMS scores and student achievement.[1]
7. In social studies/history classes (n=2), a positive, but non-significant correlation ($r = .18$, $p < .081$) appears between FPMS scores and student achievement.
8. The two measures of student achievement used in this analysis (residualized gain scores -RGS- and average stratum posttests - ASP) appear to estimate a parameter defining the same relationship between achievement measures and measures of teacher behavior.

These analyses confirmed a relationship between FPMS scores and student achievement at the elementary level. They suggested that a relationship (perhaps at a lower level) may exist between FPMS scores and student achievement at the post-elementary level. They pointed to a close relationship between FPMS scores and student task engagement at both the elementary and post-elementary levels. Task engagement related significantly with student achievement at all levels.

These results also suggested that both the ASP and RGS methods for controlling non-equivalence among groups produced quite similar results when considered over a large sample. This supported two contentions.

1. They both attempted to estimate the same population parameter.
2. Estimates produced by estimators attempting to estimate the same population parameter will converge as sample size increases (Andrews et. al, 1972; Gastwirth, 1971).

---

[1]In this document, the .05 level of significance will be accepted as a minimum level. Relationships at or above the alpha .10 level will be considered worthy of further investigation, and mentioned.

Section II
Guidelines for Conducting Predictive Validity Studies of Measured Teacher Performance

The purpose of this paper is to list the major components of study designs as they apply to estimating the predictive validity of teacher evaluation systems. It is intended to serve as a guide for preparing studies, reviewing proposed studies and conducting studies. The criteria described herein are based upon standard practices for conducting field studies (Kerlinger, 1967).

In most cases, predictive validity studies in teacher evaluation are conducted in regular classrooms, or other educational settings, where teacher performance is the independent variable and measures of student outcome are the dependent variables.

Suggested standards, such as the number of classrooms to be observed and the number of observations to be conducted, are intended only as general guidelines. These standards must be tempered by proposed designs and available resources. An estimate of cost of conducting predictive validity studies may be found in Appendix B .

Selection of the subject in which a study is to be conducted is the first matter to be decided. Unlike reliability and norming studies in teacher evaluation, where a cross section of subject areas is included in each study, predictive validity studies are subject specific. This causes predictive validity research on teacher evaluation systems to be costly since studies should be conducted in each of the subject areas in which a system is intended to be used, or until the obtained results are found to be consistent across a representative sample of subjects. First then, a subject area is selected such as reading in the elementary school, or American history in the junior high school or physics in the senior high school. Subjects may, of course, be studied simultaneously, but data analyses may be conducted separately by subjects since the measures of student outcome will be different from subject to subject, as will the context of each subject.

Selecting subjects where learning the content is least affected by forces external to the classroom, and selecting subjects for which reasonably precise measures of student outcomes are available will probably yield the most positive results in predictive validity studies.

Beyond the selection of subject(s), several decisions must be made in designing predictive validity studies. They are discussed under the general categories of sample, measures, data collection, outcomes, and analysis.

Sample

To generalize results of predictive validity studies, research must be conducted in natural teaching situations. This precludes most experimental controls such as random selection, assignment and special treatments. The sample must be representative of the population to which the findings are to be generalized. As a matter of practicality, many of these studies are school district oriented rather than state or nationally oriented. If the district has, for example, forty teachers of a commonly taught course in second grade mathematics, all of the teachers could be studied. If there are one hundred and twenty such teachers, a sample could be selected.

Sample selection must control for variations in teaching context, particularly the socio/economic status (SES) of the students (Brophy and Evertson, 1976; Veldman and Sanford, 1984). A full range of SES should be included. Occupation of the head-of-household is a reasonably accurate method of determining student SES. If those data are not available, free and reduced lunch-program data are an alternative although probably not as accurate a measure as head-of-house occupation. Geographic stratification will help to control for SES, rural/urban and other content variations.

Sample size is closely related to the anticipated unit of analysis. If the class is used as the unit of analysis , thirty teachers(classrooms) in each group is minimumly adequate. The class is a stable, generally accepted unit of analysis. It is wise to start with more if possible in the event of attrition. If inordinate attrition occurs, the classes (teachers) that dropped out of the study must be examined for possible effects on the sample.

Additional concerns include checking to see that class periods are the same length across classes and that the teachers and students are stable and not moving or dropping out of the classes during the period of study.

Measures

The independent measure in these studies is teacher performance and the dependent measure is student performance. In most studies there will be pre and post tests of students, process measures of teacher performance, and data collected on appropriate frame factors of both teachers and students (Micceri, 1984a).

Pupil pretests or covariates (Frame Factors) may include achievement, aptitude, SES and various student characteristics. These data must be collected accurately and there must be evidence that they in some way relate to study purposes. The purpose of pretests and covariates in studies where laboratory controls are not practical is to provide a means of adjusting scores to account for differences in posttest measures attributed to factors other than teacher performance. Use pretests when students have prior knowledge of a subject. Use an ability measure when content is foreign to the students. Class means may be used as estimates of variables such as SES. However, individual pupil estimates of SES better control confounding effects than class mean estimates. Keep the number of covariates to a minimum due to the sacrifice of degrees of freedom in multivariate analyses or use residualized gain scores to alleviate this problem (Klitgaard, 1978; Baldwin, Medley, and MacDougall, 1983).

A number of teacher characteristics are of interest to determine their possible relationships to variations in outcome measures. Age, experience, degrees , race and sex are commonly tested teacher characteristics. Knowledge of subject is another measure of interest. However, teacher characteristics should only be included in the design if there is a rationale for their inclusion.

One independent variable is the measure taken of teacher performance. Another may be a mediating measure of student outcome such as student engaged time. The measures of teacher performance, which are usually made by teacher observation instruments, should be content valid and meet reliability standards prior to use in predictive validity studies.

Data Collection

Several issues that must be addressed in the study design relate to how data are collected. They include the study length, the number and length of classroom observations, the number and quality of observers, and control of the content taught during the study period.

Study length may vary from short-term (two weeks) to long-term (academic year) depending upon the purpose. The key to determining length is to estimate the time required to attain a desired quantity of student growth. For example, a study designed to measure student memory of a limited number of predetermined facts may be conducted in a short-term study, while attempts to study the development of problem solving skills may require a longer period.

The number and length of observations of teacher performance is determined by what it takes to generate a stable sample of the behaviors being measured. If prior reliability data on the instrument and observers show adequate stability, then the number of observations required to collect a representative sample of teacher performance is the number at which average observation scores stabilize. As a minimum, one observation should be conducted near the beginning and one near the end of the period being studied. In general, the class period is the length of each observation, although some systems recommend various coding or rating intervals during the class period in lieu of continuous coding or rating. Entire class periods probably should be coded or rated, however, so that beginning, middle and ending performance is recorded.

The observers who record teacher performance must be trained and tested and frequently updated in the use of the teacher evaluation instrument. They should have no stake in the outcome of the study in order to protect observations from bias (Medley, Coker, and Soar, 1984). A minimum of two observers working independently should be randomly assigned to each teacher. Attempt to maximize the number of observations.

Select the content to be taught. Specify and, to the degree possible, monitor the content to be presented during specified periods.

Outcomes

What students know and can do as a result of schooling are the products that researchers attempt to quantify and use as dependent variables in teacher effectiveness/evaluation studies. The common variable has been student achievement. However, because of the multitude of intervening variables over which teachers have no control that may influence achievement scores, researchers have for the past few years been using mediating measures of student outcome as a supplement to achievement measures to reduce the effects of non-teacher influences on outcome measures. *Another approach to this problem utilizes measures of student affect to evaluate both teacher affect and effect.*

Measures of outcome, then, *may* be considered in *three* dimensions - - post tests of knowledge and performance and mediating tests that are conducted at the time the teachers performance is being observed. *The third could include measures of pupil affect, eg.--motivation, curiosity, satisfaction.*

Achievement tests that have been examined for their validity and reliability and standardized , at least over the population in the study, are the most common tools used as posttests in these studies. It is vital that the content taught match the content tested in the classes observed to maximize the teachers' influence on the outcome measure. Tests must also have a high ceiling in order to assess the full range of student learning and not restrict variability among classes.

If the subject matter is foreign to students, such as learning Russian as a foreign language, an ability measure should be used as a pretest. If the students have prior knowledge of the subject, which is often the case, use a pretest that measures knowledge of the specific content that is to be taught during the study period. Both pre and posttests should vary in cognitive level. Non-cognitive measures should be used when available. Much of this depends, of course, upon the purposes of the study.

Timing of test administration is an important issue. Since regular classrooms are the laboratories for the majority of these studies and budgets are frequently minimal, often researchers must use the tests given in the normal course of events within the school district. This sometimes means using general measures of achievement that frequently are given during the spring of the preceding year as pretests. This practice is questionable for at least three reasons. There is serious question of content validity of these exams when a test given in the spring of the year is used as a pretest for the fall of the next year, variations in retention between SES levels across time add error to estimates of growth between pre and post measures, and high levels of student attrition may substantially reduce class sizes. It is best practice to pre and post test as close to the study period as possible.

The use of covariates (also discussed under measures in this paper) helps to partial out variations in outcomes attributed to sources other than teacher performance. SES is most commonly used as one of the most valuable covariates since it frequently relates to prior achievement. Free lunch data may provide an estimate of SES. It is preferable to code SES using parent income or occupation at the individual student level when possible. The class mean of these scores may be used as an SES variable, however, individual pupil estimates of SES better control confounding effects. Other potential covariates include class size and classroom conditions such as

lighting and temperature control. Keep the number of covariates to a minimum since they reduce the number of degrees of freedom and the potential of obtaining significant results.

Non cognitive measures should also be explored as measures of student outcome. Some studies have used measures of affect. One such measure is student response to questions concerning their satisfaction with school. In the future, researchers should attempt to study success in life of various cohorts of students that have been trained by various groups of teachers.

Mediating measures of student outcome show great promise in teacher effectiveness/ evaluation studies. These measures are taken at the same time measures of teacher performance are generally taken, thereby, obtaining a measure of the immediate effect of the teacher's performance on the students. The most common of these measures has been student task engagement. Even though the literature shows a curvalinear relationship between task engagement and achievement, the studies conducted to date, using it as a mediating variable, have consistently found a positive and significant relationship to achievement.

Methods of collecting student task engagement vary across studies from coding individual student activity to class estimates of the number of students not on task during specified intervals. With the results for both methods being in the same direction and being significantly related to outcomes, either method of data collection seems adequate.

Completed homework assignments, providing they are closely controlled, student characteristics and home characteristics may also be considered as potential mediating variables.

Statistical Analysis

Independent variables include classroom performance scores or evaluation scores and scores on mediating variables (mediating measures may be used both as independent and dependent variables in these studies). Analyses may be conducted at either the school or classroom level. This depends, of course, upon whether or not there are multiple teachers participating in the study in each school.

Data should be analyzed by cognitive level and by total score on the dependent variables. These analyses should use residualized pupil gain scores as the dependent variable. Through the use of multiple regression these scores attempt to adjust for heterogeneity of pupil abilities. Pretests and/or covariates are regressed to individual outcome scores. A predicted score for each

student is computed using the resulting regression model, and obtained outcome scores are subtracted from predicted scores. The class mean of individual residualized gain scores may then be used as a dependent variable and correlated with the desired independent variable (Klitgaard, 1978).

Combined Studies Analyses

The most powerful predictor of student learning is what the students know prior to entering the class. What the teacher does is the second most powerful predictor. Up to eighty percent of the variance in outcome scores is accounted for by the students' prior knowledge. Thus, a maximum of about twenty percent of the variance can be attributed to the teacher's instruction. This characteristic has consistently led to low relationships between measures of teacher performance and student outcomes.

Because validity studies typically include fewer than thirty classrooms, and teacher behaviors at best account for only a small percentage of pupil learning, statistically significant effects at the generally accepted .05 level are rare. If several studies with sample of less than thirty classrooms are conducted, and then pooled, the sample size can become large enough to detect moderate size effects. This procedure is generally termed meta analysis (Rosenthal, 1978). Research consistently suggests that teacher effects on achievement tests will rarely be of a magnitude sufficient to be defined as significant for sample sizes smaller that thirty to one hundred classrooms. Thus, pooling more than one study in more than one subject area is more likely to identify the moderate pupil changes that result directly from effective teaching. Therefore, the long range goal in predictive validity research is to conduct a number of studies that represent at least the general levels and subjects of interest and to conduct meta analyses of the pooled results.

OUTLINE

Guidelines for Conducting Predictive Validity Studies of Measured Teacher Performance

This outline summarizes the major components of predictive validity study designs for teacher evaluation systems from Section II, Chapter 4 of the 1985/1986 Teacher Evaluation Study. Criteria for these studies are based upon standard practices for conducting field studies where teacher performance is the independent variable and student outcome the dependent variable.

Suggested standards are intended only as guidelines. Selection of the subject(s) to be studied is the first decision. Following selection of subject(s), designs for predictive validity studies may be discussed under the categories of sampling, measures, data collection, outcomes and analysis.

Sampling

I. Predictive validity studies are conducted in natural teaching situations

II. The sample must be representative of the population to which the findings are to be generalized

III. Sample selection must control for variation in teaching context
   - A. A full range of SES should be included
   - B. Geographic stratification will help to control for SES, rural/urban, and content variations

IV. Sample size is closely related to the anticipated unit of analysis
   - A. The class is a stable, generally accepted unit of analysis
   - B. Thirty teachers (classrooms) in each group is minimally adequate
   - C. Start with more teachers (classrooms) if possible to guard against attrition
   - D. Check to see that class periods are the same length and students are stable

Measures

I. The independent measure is teacher performance and the dependent measure is student performance

II. Most studies will have pre and post tests of students, process measures of teacher performance, and frame factor data on teachers and students

III. Covariates (Frame Factors) may include achievement, aptitude, SES, and student characteristics

   A. Use pretests when students have prior knowledge of a subject

   B. Use ability measures when content is foreign to students

   C. Class means may be used as estimates of variables such as SES

IV. Teacher characteristics such as age, experience, and sex may be of interest to determine their relationship to either teacher or student outcomes

V. Student engaged time may be used as a mediating measure as either an independent or dependent variable

VI. All measures should be content valid and meet reliability standards

Data Collection

I. Study length may vary from short-term to long-term depending upon the purpose

   A. The key is to estimate the time required to attain desired student growth

II. The number and length of teacher observations is determined by obtaining a stable and representative measure of performance

   A. As a minimum, one observation should be conducted near the beginning and one near the end of the study period

   B. Observation length is generally a full class period

III. Observers must be trained and tested in the use of the evaluation instrument

   A. Observers should have no stake in the outcome of the study

   B. Randomly assign a minimum of two observers to each teacher

   C. Maximize the number of observations

IV. Specify and, to the degree possible, monitor the content presented during study periods

Outcomes

I. The most commonly used outcome variable is student achievement
   A. Content taught must match test content
   B. Tests must have a high ceiling to assess the full range of student learning

II. Mediating measures of student outcome serve as a supplement to achievement
   A. Measures such as student affect may be used to measure both teacher affect and effect

III. Use an ability measure as a pretest if the content is foreign to students and a content specific achievement test if students have prior knowledge of the content
   A. Items on pre and post tests should vary in cognitive level
   B. Non-cognitive measures should be used when available

IV. Timing of test administration is important
   A. Give pre and post tests as close to the study period as possible
   B. Often tests must be given in the normal course of events within school districts
   C. Exams given in the spring and used the next fall as pretests raise serious questions of validity

V. Covariates may help to partial out variations in outcomes attributed to sources other than teacher performance
   A. SES is one of the most frequently used covariates
   B. Other covariates include class size and conditions
   C. Keep the number of covariates to a minimum since they reduce the number of degrees of freedom

VI. Non cognitive measures may also be explored as measures of student outcome
   A. Student response to questions regarding satisfaction is one such measure

VII. Mediating measures of student outcome show promise in teacher performance/ evaluation studies
   A. Student task engagement is the most common of these measures
   B. Completed homework assignments may also be considered as a potential mediating variable

Data Analyses

I. Analyses may be conducted at the school, classroom, or individual level, depending upon the amount of data gathered for each individual student and teacher, and study intent.

II. Most studies are based on correlations between independent (teacher) measures and dependent (student) measures

III. Residualized gain scores are recommended to control for preexisting differences among classes

   A. Residualized scores allow the final analysis to be a correlation between teacher and student measures

Combined Studies Analyses

I. The long range goal in predictive validity research is to conduct a number of studies that represent at least the general levels and subjects of interest, and to conduct meta analyses of the pooled results

   A. If several studies are combined, the sample size can become large enough to detect moderate size effects

   B. Teacher effects on achievement are rarely of a magnitude sufficient to be defined as significant for sample sizes smaller than thirty to one hundred classrooms

   C. Pooling more than one study in more than one subject is more likely to identify moderate pupil changes resulting directly from effective teaching

Section III

An Exploration of Potential for Conducting Macro Predictive Validity Studies on the Quality of Education

The major purpose for conducting large scale studies (macro studies) is to evaluate the quality of education among schools and districts and to monitor possible effects that special programs, such as the Beginning Teacher, Master Teacher and Career Achievement programs, may have on the quality of education across the state. These studies compare variations in teacher performance scores in to variations in a variety of student outcome measures. Should higher scores on teacher observation instruments across the state associate with higher standardized achievement results, such comparisons would provide empirical support for the effectiveness of the educational system.

These studies can be conducted at various levels:

1. Class or teacher level, relating class mean scores (usually achievement) with teacher performance scores either within or across districts.
2. School level, relating school mean achievement scores with teacher performance scores either within or across districts.
3. District level, relating district mean achievement scores with district wide teacher performance scores.

For all the suggested levels, separate analyses may be conducted at the elementary, middle and secordary school levels. Analyses can also be conducted by subjects and a number of other frame factors.

Most of the research tenets detailed in Section II of this chapter also apply to providing guidelines for macro predictive validity studies. However, macro studies present unique problems and advantages. For example, the larger and longer term macro studies allow for greater generalizeability but lose considerable control of relevant variables. Specific studies within schools or classes allow for considerable control, but cannot easily be generalized to

broader questions. Calling upon the power of large numbers, one may assume that despite all the possible confounds, if sufficient subjects are used in a macro study, such confounding factors should tend to even out.

Two primary assumptions underly macro analyses:

1. The sample of teachers studied is representative of the population to which the findings are to be generalized.
2. The general effects of teacher performance will vary adequately and associate with differential performance by students.

Current Impediments to Macro Studies

A preliminary search was conducted during 1985-1986 to determine the feasibility of conducting macro predictive validity studies of the nature cited above. The largest common pool of teacher performance scores at this time was the FPMS Master Teacher Program performance scores. An available measure of student outcome was the Comprehensive Assessment Program (CAP).

These measures proved to be inadequate for purposes of conducting macro studies for the following reasons:

1. Inadequate representation of teachers in most schools,
2. Inadequate representation of pupils in some schools,
3. Non-random selection of teachers,
4. High teacher performance scores and smaller variability than generally occurs with the FPMS instrument,
5. Non-random selection of students,
6. Inadequate representation of grades on the CAP (complete data were available only for grades four to six), and
7. The existence of clusters of gifted students in a few schools.

The conduct of macro predictive validity studies is fundamental to evaluating the quality of education by subject, level, school and district. To do so, however, will require commonly used, validated measures of teacher performance and student outcomes.

CHAPTER 5

Selected Analyses of 1984-1985 MTP Data

This chapter includes six analyses of data from the performance observations conducted during the 1984-1985 Master Teacher Program (MTP).

These analyses are presented in the following sections:

1. Study of the relationships between Florida Performance Measurement System (FPMS) or Teacher Assessment and Development System (TADS) scores and special instrument scores. This is an examination of the validity of the special instruments to impart their value in adding unique performance information to MTP evaluations. Significant correlations may lend support to the concurrent validity of the instruments, while excessively high correlations would suggest that the special instruments add no unique information to that obtained from the FPMS or TADS instruments.

2. An examination of score trends over time on each of the instruments used in the MTP observations. Assuming no systematic relationship between applicant ability and date of observation, variations in mean scores over time would suggest contaminating factors operating to influence scores on the observations.

3. A study of the differences in scores obtained on the FPMS Summative observations between laboratory and non-laboratory settings. The need for this analysis was identified after MTP observations were conducted and applicants expressed the opinion that observations conducted during labs were not accurate reflections of their teaching performance.

4. An examination of the distribution shapes and point ranges by quartile for each instrument used in the MTP. Heavily skewed distributions indicate ceiling or floor effects and bring in to question the value of the instument in discriminating between levels of performance. The number of points of discrimination by quartile is of importance in setting standards. If the error factor is equal to the number of points of discrimination within a quartile, candidate selection may be attributed to chance rather than performance.

5. An analysis of FPMS and TADS factors. Factor analyses reveal factor unidimensionality and internal consistency data. These data estimate the degree to which sub-scores are either related or independent.

6. A study of the number of observers used in the 1984-1985 MTP in each county and the number of observations made by each observer. The value of this analysis is in estimating the number of observers required to conduct a program of this nature and in describing the use of trained observers in the 1984-1985 administration of the M.T.P.

The data comprising the sample for these analyses consisted of all MTP observations received prior to August 1, 1985. The following are the total samples available by instrument:

Single Observations

FPMS Summative	40,641
TADS Instrument	7,022
Student Services	3,577
Library/Media	717

Relationships Between FPMS or TADS Scores and Special Instrument Scores

Table 1 presents the intercorrelations among instrument scores. Sample sizes for the TADS correlations with the special instruments are quite small and neither of the TADS correlations differ significantly from zero. Both FPMS correlations with the special instruments are significant.

Table 2 presents inter-instrument correlations calculated within the grade level divisions of the FPMS Summative instrument. When separate correlations are calculated for the Primary/ Secondary dichotomy used in scoring the FPMS Summative, it is seen that the correlation obtained in Table 1 resulted from the K-5 group. Moreover, the magnitude of the correlation of FPMS observations and either of the other instruments is approximately equal. A higher correlation between FPMS and Student Services was obtained for the total group of observations (Table 1) because of the relatively smaller proportion of secondary level observations comprising the total (25 per cent of the Student Services observations were secondary level, in contrast with 58 per cent of the Library/Media observations).

Results from this analysis show a comparatively high but non-significant relationship between TADS and either of the two special instruments. This may be a problem of sample size. The FPMS relationship with both instruments is significant at the .01 level although the magnitude of the obtained correlation is lower than that of TADS. Since the FPMS in particular has a documented teacher effectiveness research base and has been content validated on that basis, this relationship lends credibility to the use of the special instruments in evaluating instructional personnel. The magnitude of the correlations, however, are not adequate to support the notion that the special instruments duplicate what is already measured on the FPMS or TADS instrument.

TABLE 1

Correlations Among Instruments
Master Teacher Program 1984/1985

Special Instruments	Teacher Evaluation Instruments	
	FPMS	TADS
Student Services	0.27 * (n=153)	0.59 (n= 8)
Library/ Media	0.18 * (n=318)	0.26 (n= 36)

* $p<0.01$

TABLE 2

FPMS Correlations With Other Instruments
Within FPMS Grade Level Divisions

	FPMS Performance	
Special Instruments	Primary (K-5)	Secondary (6-12)
Student Services	0.38 * (n=115)	0.07 (n= 38)
Library/ Media	0.36 * (n=132)	0.08 (n=186)

* $p<0.01$

MTP Observation Score Trends Over Time

A rough index of score trends over time is obtained by examining average scores by month of observation. Such averages are provided in Table 3.

The FPMS instrument showed an increment in average score of approximately 4 points on observations conducted in April, 1985 or later. The TADS instrument showed a similar increment which occurred one month earlier, while a smaller increment occurred in May on the Student Services instrument. The presence of this effect on the Library/Media instrument is difficult to demonstrate given the low number of observations conducted in spring and summer.

The causes of score increments across the time period studied are unknown. Avenues of inquiry in investigating these changes include: 1) the training being conducted in many districts which may have raised the quality of candidates across time, 2) competition between schools and districts which may have caused observer bias, and 3) pressures from within schools to score teachers high which again may have caused observer bias.

TABLE 3

Score Trends Over Time

FPMS Instrument

Month	N	Median	Interquartile Range
December	1102	34	12
January	8608	35	10
February	28616	35	10
March	378	35	12
April	475	39	17
May	1090	40	10
June	171	39	11
July	11	38	9

TADS Instrument

Month	N	Median	Interquartile Range
December	803	72	18
January	2820	75	14
February	3214	76	13
March	61	80	9
April	82	81	3
May	235	93	3
June	98	81	5
July	1	82	-

Student Services Instrument

Month	N	Median	Interquartile Range
December	2	23	1
January	79	27	6
February	3282	26	7
March	59	26	8
April	18	27	10
May	74	29	9
June	42	29	7
July	1	27	-

TABLE 3 (continued)

Score Trends Over Time

Library/Media Instrument

Month	N	Median	Interquartile Range
December	0	--	-
January	46	23	7
February	628	22	7
March	26	23	4
April	0	--	-
May	12	25	3
June	2	21	3
July	0	--	-

Differences in FPMS Observation Scores Between Laboratory and Non-laboratory Settings

Table 4 presents performance information on observations conducted in what may have been laboratory environments. Three criteria were used to select these observations:

1. The method was independent study.
2. The observation was conducted in a self-contained classroom.
3. The subject was Math, Science, Home Economics, or Art.

The observations conducted in the labs received the standard adjustment applied to the independent study teaching method. As evidenced in Table 4, fewer (but not significantly fewer) effective indicators are typically coded in the lab environment, and the resulting observation scores are somewhat lower than in non-lab settings. However, it is important to recognize that the data may be contaminated because of the imprecise method of identifying observations conducted in laboratories. These results do suggest that more direct study of laboratory teaching behavior as measured by the FPMS Summative is justified.

A study is currently being conducted in which science teachers are being observed in regular classroom settings and in laboratories. This study, scheduled to be completed in the Spring of 1987, should provide a more precise answer to the question of whether significant differences exist between teacher performance in classrooms vs. laboratories as measured by the FPMS.

TABLE 4

Lab Environments and FPMS Scores

	Elementary		Secondary	
	Lab	Non-lab	Lab	Non-lab
N	58	17947	318	22318
Performance Scores				
Mean	36.44	37.68	32.09	32.89
Median	36.89	38.00	30.94	33.00
Std Dev	7.88	7.03	7.23	7.09
Interquartile Range	12.00	10.00	9.98	10.00
Number of Effective Indicators Used				
Mean	14.83	16.46	13.86	16.07
Median	15.50	17.00	14.00	17.00
Std Dev	3.30	2.42	3.90	2.84
Interquartile Range	4.00	3.00	6.00	3.00

Distribution Shapes and Point Ranges for Performance Evaluation Instruments

The psychometric purpose of the performance evaluation instruments as applied in the Master Teacher Program is the discrimination of the top 25 per cent of all candidates who complete the performance evaluation process. Distribution shape plays a role in each instrument's effectiveness in accomplishing this task. Two inter-related aspects of distribution shape were examined in relation to the four instruments used in the MTP: range of points in the upper quarter of the distribution and distribution skewness.[1]

To discriminate individuals at the top end of an instrument's scale, these individuals must score over a range of scale points. The degree to which a range of scale points is available for candidates reflects the degree to which performance differences may reasonably and confidently be detected. The range of points available in any portion of a distribution is in part attributable to the absolute range of the instrument's scale (e.g. the Media instrument has a total scale length of 25 points while the TADS instrument has a scale length of 82 points) and in part to the distribution's skewness. For discrimination at the top of the distribution, negative skewness is potentially problematic. Negative skewness reflects a tendency for most candidates to score near the maximum score of the instrument (ceiling effect), with a few candidates scoring toward the low end of the scale.

Table 5 presents descriptive data on each instrument used in performance assessment for the Master Teacher Program (because the FPMS instrument provides separate distributions for primary and secondary academic levels, these distributions are evaluated independently). Histograms of each of these distributions are reproduced in Figures 1 through 5.

Comparison of these five distributions shows a clear dichotomy of classes resulting from differences in underlying scale construction. Both the FPMS and Student Services instruments' scales are normatively derived from empirical item frequencies. That is, the scale score for an individual's performance depends upon his performance relative to other Master Teacher candidates' performances. The distributions of the FPMS instrument and the Student Services

---

1. Measures of kurtosis may also contribute to discriminability in some distributions. In the distributions obtained from the Master Teacher observations, the role is minor compared to scale range and skewness.

instrument appear relatively normal in shape. In contrast, the Library/Media and TADS instruments' scales are criterion derived, each item being scored according to observer judgement of the quality of the observed performance required by the item. The scale score is derived without regard to the performance of other applicants observed using these systems. Each of these distributions shows a clear ceiling effect with corresponding negative skewness.

The effect of these differences in classes of instruments on discrimination ability in the upper quarter is evidenced by the range of scale comprising the top 25 per cent of each distribution. The FPMS distributions provide 14 points and 15.5 points of discrimination in the upper quarter for the primary and secondary levels respectively (this represents 23 per cent and 26 per cent of the theoretical range of the FPMS instrument at these two levels). Similarly, the Student Services instrument provides seven scale points of discrimination in the upper quarter (19 per cent of the total range). In contrast, the TADS instrument provides only 3.5 scale points above the 75th percentile (4 per cent of the total range), while the Library/Media instrument distribution provides 1.5 points in this range (6 per cent of its range).

The practical importance of these statistics is that most of the scale ranges on the TADS and Library/Media instruments are functioning to discriminate individuals at the low end of the distribution. For minimal competency testing, where the purpose is to discriminate at the scale's low end only, this may be desirable. However, this property is at odds with the psychometric task required of the Master Teacher Program performance evaluation.

TABLE 5

Instrument Distributions Obtained in the Master Teacher Program

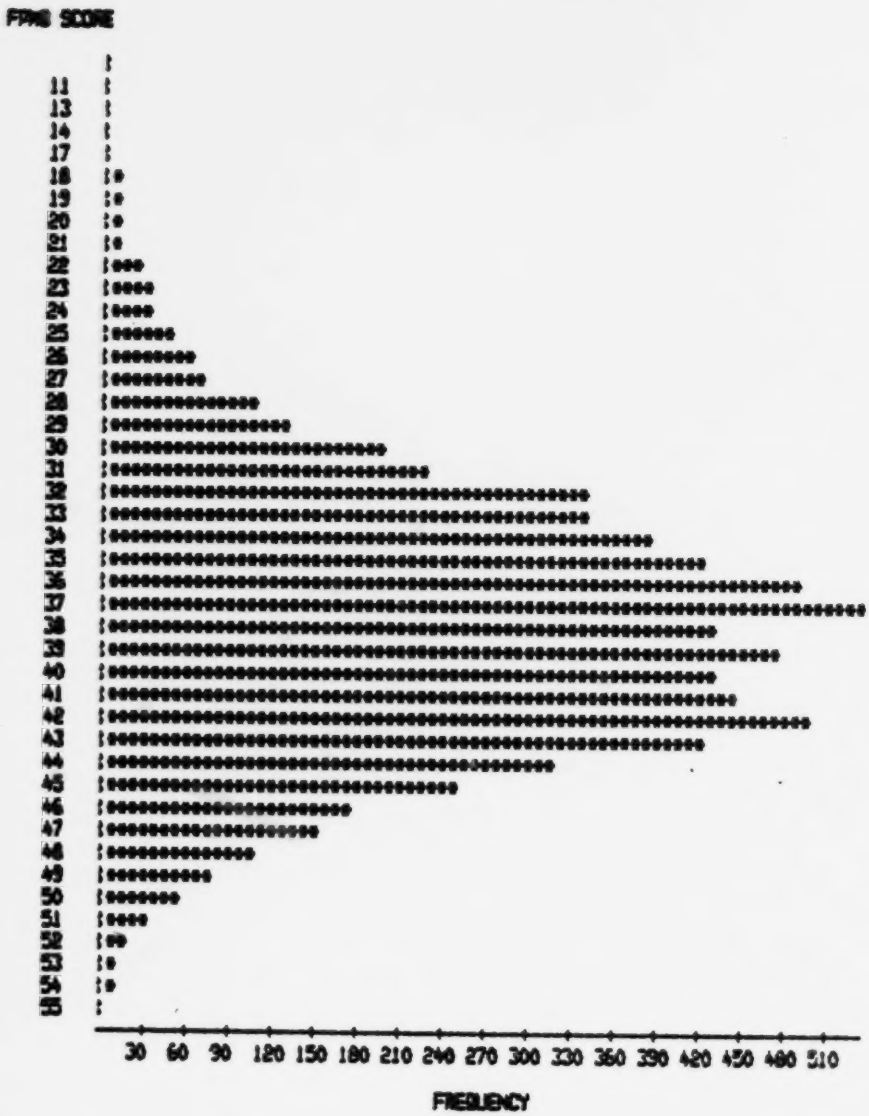
Characteristic	Instrument Distribution				
	FPMS Elem	FPMS Sec.	TADS	Student Services	Media
Theoretical Range	60.9	59.9	82.0	36.0	25.0
Obtained Range	42.0	41.0	57.5	31.5	21.5
25th Percentile	33.5	29.0	66.0	23.5	17.0
75th Percentile	41.0	36.5	78.5	29.0	23.5
Lowest Qtr (Range)	20.5	18.0	41.5	19.0	13.5
Mid-Spread	7.5	7.5	12.5	5.5	6.5
Highest Qtr (Range)	14.0	15.5	3.5	7.0	1.5
Mean	37.2	32.8	70.9	26.1	19.8
Median	37.5	33.0	79.8	26.0	21.5
Mode	38.5	33.0	82.0	29.0	25.0
Std Deviation	5.6	5.7	9.8	3.9	5.0
Skewness	- 0.2	- 0.1	- 1.4	- 0.3	- 1.2
Kurtosis	0.1	- 0.1	1.9	0.4	0.9

FIGURE 1

MTP DATA 1985

DISTRIBUTION OF FAMS OBSERVATION SCORES
LEVEL=Elementary

FREQUENCY BAR CHART



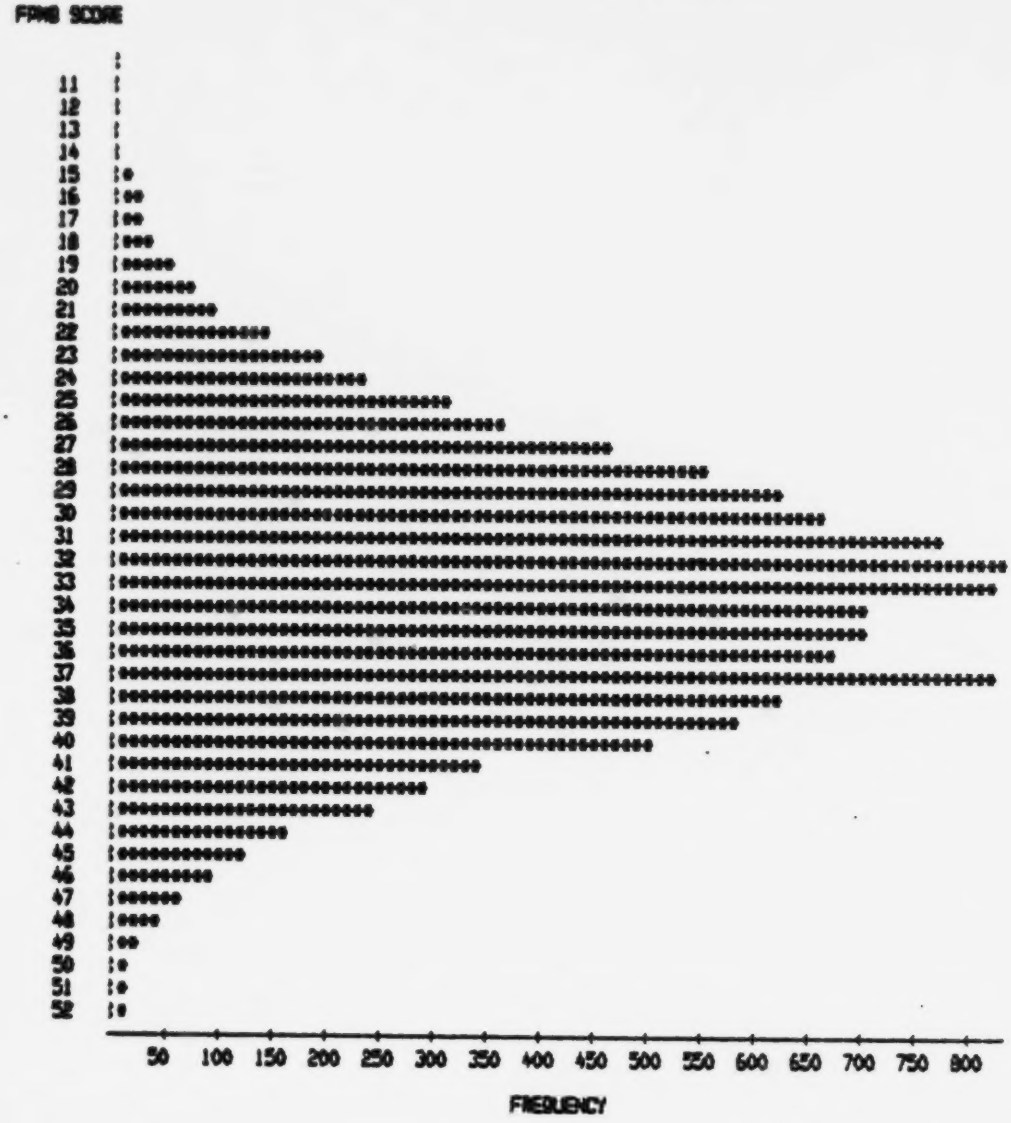
BEST COPY AVAILABLE

FIGURE 2

MTP DATA 1985

DISTRIBUTION OF FFMS OBSERVATION SCORES
LEVEL=Secondary

FREQUENCY BAR CHART



FIGURE

MTP DATA 1985

DISTRIBUTION OF MEDIA OBSERVATION SCORES

FREQUENCY BAR CHART

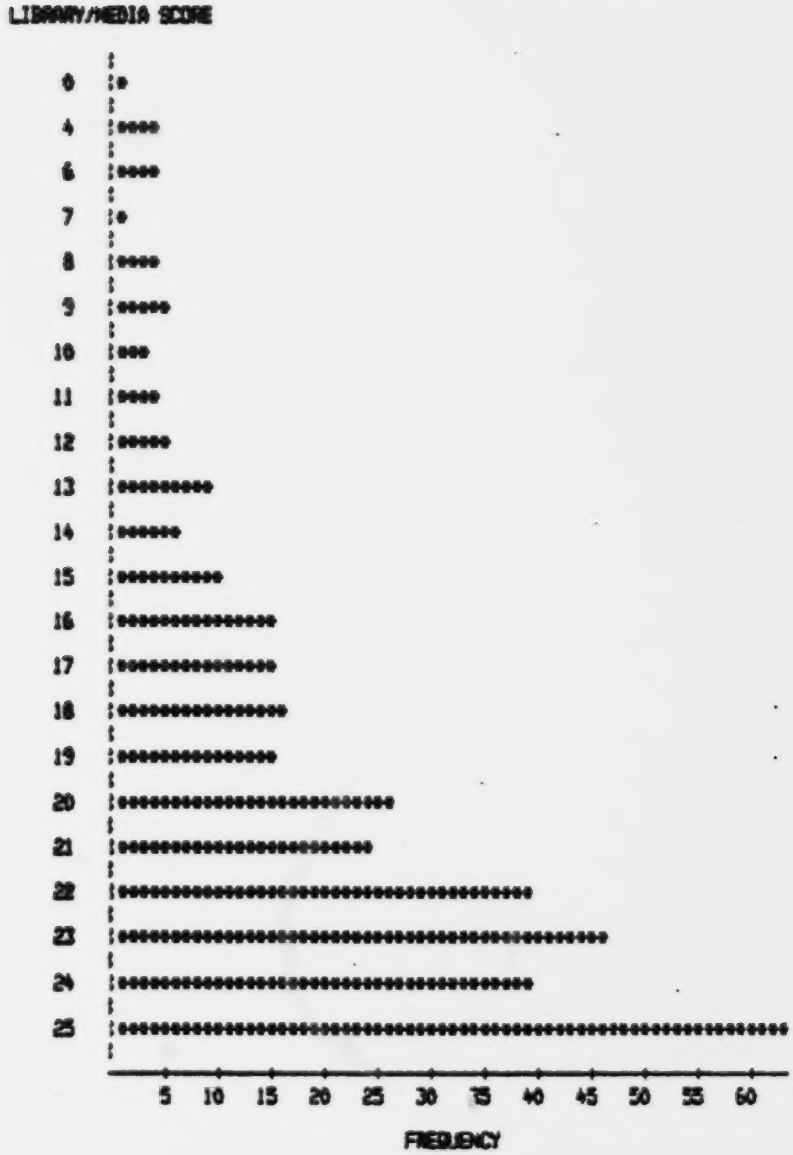


FIGURE 4

MTP DATA 1985

DISTRIBUTION OF STUDENT SERVICES OBSERVATION SCORES

FREQUENCY BAR CHART

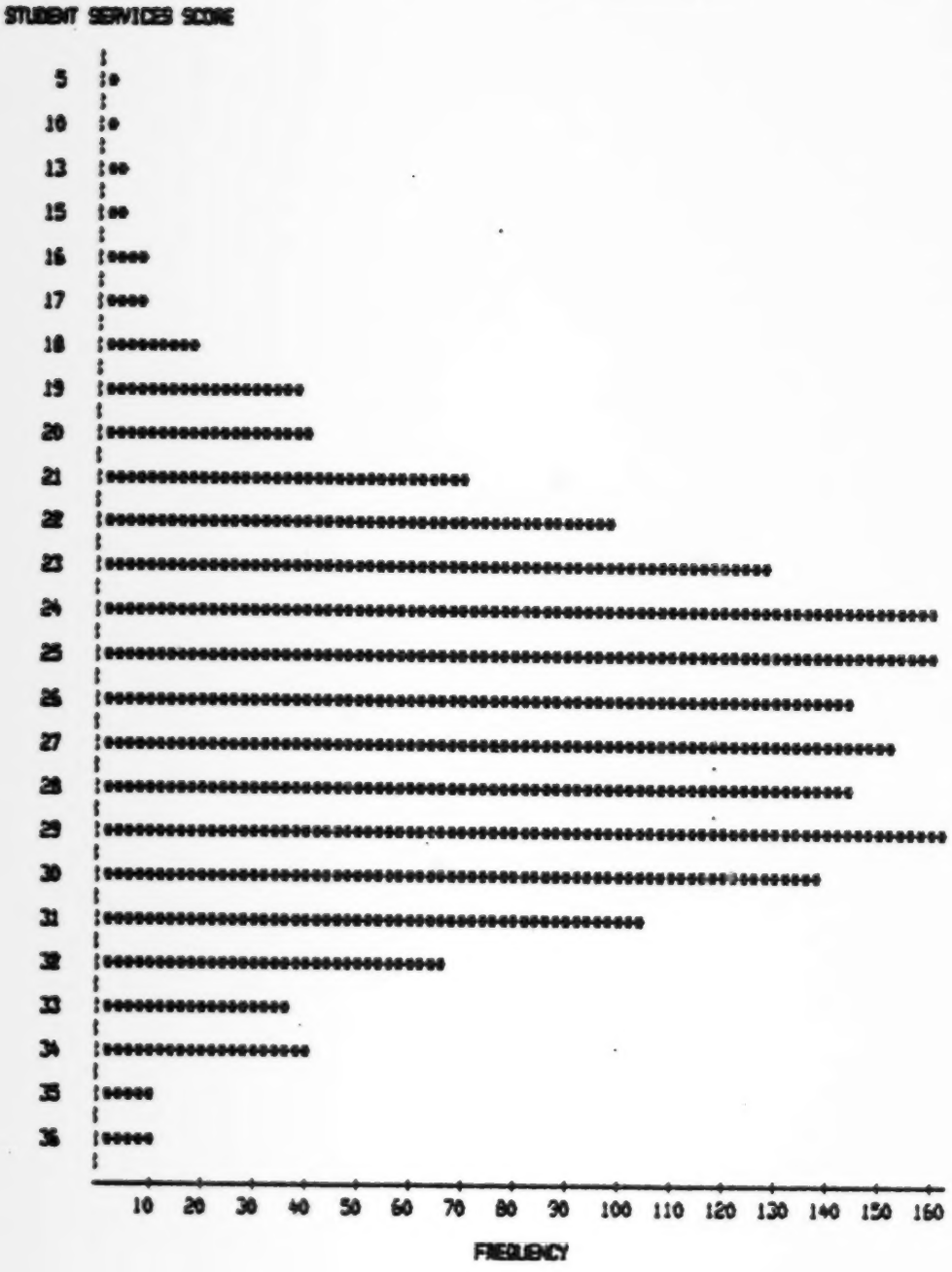


FIGURE 5

MTP DATA 1985

DISTRIBUTION OF TADS OBSERVATION SCORES

FREQUENCY BAR CHART

Factor Analysis of FPMS and TADS Instruments

Factor analyses were conducted on the two major instruments in the MTP to more fully understand the meaning of the total scores for each instrument. These analyses provide information for interpreting inter-item correlations and instruments dimensionality. Additionally, factor analyses help to confirm underlying assumptions of the instrument's construction such as the existence of domains and effective versus ineffective items on an instrument.

Factor analysis is a statistical technique of reducing a matrix of correlations to more easily understood dimensions. As applied to instrument development, factor analysis is used to aid in grouping items and to describe more fully the attributes being measured.

Variability of scores on items may be divided into two components: that variability which is shared with other items and variability unique to each item. A third score component is error variability. Knowledge of this component and its statistical treatment is not essential to understanding the analysis presented here and it will not be dealt with. Factor analysis partitions shared variance into a small number of hypothetical constructs (factors). In the example of the FPMS instrument analysis presented here, the 39 items of the instruments were conceptually reduced to four factors.

Associated with each factor is an eigenvalue or characteristic root. This is a number which represents the amount of total variance accounted for by the factor. In contrast, the "proportion of variance" reported in this analysis (e.g. Table 6) represents the proportion of common variance explained by each factor. The mathematics underlying factor analysis insures that maximum variance is accounted for by each factor as it is extracted, and that each factor accounts for variance independently of all other factors. That is, the first factor accounts for more variance than any other factor which may have been extracted, the second factor accounts for more of the remaining variance (adjusting the correlation matrix for the first factor) than any other factor and is uncorrelated with the first factor, etc.

The factors as initially extracted typically have ambiguous interpretations. Usually, all variables correlate to some degree with each factor. Because the purpose of factor analysis (as presented here) is "understanding", the factors are rotated to reduce ambiguity of interpretations. (If a small number of factors have been obtained from a large correlation matrix, but they make no

logical sense, then the procedure has accomplished nothing in lending understanding of what is being measured.) The term "rotation" refers to a geometric conceptualization of factors. In essence, the factors are shifted to achieve a more interpretable pattern of correlations with variables. Two general classes of rotation are used. Orthogonal rotation, as the term implies, results in factors which continue to be uncorrelated with each other. Oblique rotation results in factors which are allowed to correlate to the degree necessary to maximize factor interpretability. The interpretability sought is called "simple structure", which means that the final factor positions are selected so that most variables correlate highly with one and only one factor.

Factors are interpreted by examining correlations of factors and the original variables (in this case, items) in the correlation matrix. The meaning of items which correlate highly with a factor are combined conceptually to yield a meaning for that factor. When factors are allowed to correlate (oblique solutions), two types of factor-item correlations are legitimate considerations: zero-order and partial correlations. Zero-order correlations are simple correlations between factors and items, while the partial correlations are correlations between factors and items after statistically controlling (partialing) for the other factors. As an analogy, consider three measures: student achievement scores, intelligence test scores, and motivation assessment scores. One may interpret the correlation between achievement and intelligence (zero-order correlation), or the relationship between achievement and intelligence after controlling for motivation (partial correlation).

Final estimates of variance explained by each factor are obtained after the number of factors has been established and the inter-factor correlations have been determined. When factors are correlated, two estimates of explained variance are provided, roughly analogous to zero-order and partial correlations. One estimate is of explained variance ignoring other factors, and the other is explained variance after eliminating variance in common with all other retained factors in the solution. These estimates of explained variance are used to interpret the relative importance of the factors.

Factoring Procedures Used in this Analysis

The FPMS and TADS instruments were factored using the following procedure:

1. Four criteria were used to determine the number of factors to be retained: a) examination of scree plots, b) comparison of magnitudes of obtained eigenvalues with estimated

eigenvalues from a matrix of random correlations, c) cumulative proportion of variance accounted for by factors, and d) interpretability of factors.

2. Squared multiple correlation coefficients were used as initial communality estimates in the matrices to be factored.
3. Both varimax and oblique rotations were performed. The magnitude of the inter-factor correlations (in the oblique rotation) dictated the final structure reported here. If the typical inter-factor correlation was at least .30, the oblique solution was selected.

FPMS Summative Instrument

Separate analyses were conducted on the elementary and post-elementary groups as defined in the Master Teacher Program. Because the factor structures which emerged from these analyses were virtually identical, and also to reduce redundancy in this report, a single description of the results is provided.

For each group, four factors were retained and were rotated to oblique solutions. Tables 6 and 7 present the obtained eigenvalues and inter-factor correlations respectively for the post-elementary group. Tables 11 and 12 contain the same information for the elementary group.

TABLE 6

Eigenvalues of Four Factors Retained
FPMS Summative Instrument
Post-Elementary Level

Factor	Eigenvalue	Proportion of Variance	Cumulative Proportion
1	4.697	0.54	0.54
2	2.245	0.26	0.80
3	1.156	0.13	0.93
4	0.986	0.11	1.04

TABLE 7

Factor Correlations
FPMS Summative Instrument
Post-Elementary Level

	Factor 2	Factor 3	Factor 4
Factor 1	.31	.44	.29
Factor 2		.37	.07
Factor 3			.31

Table 8 presents factor loadings for the FPMS post-elementary group. The first factor includes items indicative of conducting classes in a business-like, task-oriented, efficient manner. Highest loadings are on items of providing seatwork, circulating and assisting students, and providing practice. Additional items loading on this factor are maintaining focus and momentum, handling materials in an orderly manner, and conducting reviews. Teacher affect items (expressing enthusiasm and body language) also load on this factor. The correlations between affect items and items indicative of task-oriented, efficient teaching is an interesting phenomenon that merits further investigation. The implications are that "on-task" teachers are also affectively more effective -- one group of behaviors may support the other in measures of teacher effectiveness.

The second factor clearly loads the majority of ineffective indicators, which provides support for the use of this side of the instrument as a separate scale.

Factor three loads items measuring the teacher's treatment of content (Domain 4.0), including higher order questions and the emphasis of important points.

Factor four deals with verbal interaction between teacher and students. This factor loads on questioning and the provision of feedback.

Table 9 presents the zero-order correlations between factors and items for this analysis. The analogous loadings and correlations for the elementary group are presented in Tables 13 and 14.

Table 10 presents the final estimates of variance explained by each of the four rotated factors. Variance estimates both ignoring other factors and controling for the other factors are provided. Factor two explains the greatest proportion of unique variance, and is closely followed by Factor one. The importance of the ineffective behaviors factor may be population-specific outcome -- these behaviors were rarely evidenced in the Master Teacher observations, and when present made important contributions to scores. A comparison of the importance of this factor with an analysis from other teacher observation data would reveal whether this is only an artifact of the MTP population. The final estimates of explained variance for the Elementary level candidates is provided in Table 15. The relative importance of factors (in terms of explained variance) is identical to that of the post-elementary group, and the magnitude of the variance explained by each is comparable.

In summary, the factor structure of the FPMS instrument is supportive of the theory underlying its construction. Effective and ineffective behaviors occur relatively independently of each other and some of the domain demarcations are supported by the factor structure (e.g. Domain 4.0). The appearance of a verbal interaction factor (Factor four) suggests that teachers' questioning behaviors operate relatively independently of other behaviors assessed on the instrument. Exploratory work with an interaction subscale is recommended. The similarity of the factor solutions across elementary and post elementary levels lends additional credence to the stability of the factor structure of the FPMS instrument. The identical four factors emerged from each set of observations and they retained approximately equal importance and approximately equal factor inter-correlations.

TABLE 8

Reference Structure
Semi-partial Correlations
FPMS Summative Instrument
Post-Elementary Level

	Factor			
Item	1	2	3	4
Begins Instruction Promptly				
Handles Material Orderly	.46			
Orients/Maintains Focus	.46			
Beginning/Ending Review	.32			
Single Factual Questions				.57
Analysis/Reasoning			.32	.34
Corrective Feedback				.58
Specific Academic Praise				
Provide for Practice	.52			
Homework/Seatwork Assign	.48			
Circulates and Assists	.56			
Concept Treatment			.44	
Discuss Cause/Effect			.58	
State/Apply Academic Rule			.29	
Value Judgement			.50	
Emphasize Important Points			.33	
Express Enthusiasm/Challenge	.39			
Body Behavior Shows Interest	.37			
Stops Misconduct				
Maintain Momentum	.35			
Delays		.37		
Not Organized		.40		
Talk Unrelated to Topic		.45		
Multiple Questions,Unison Resp.				-.33
Non-academic Questions		.27		
Ignoring,Sarcasm,Disgust		.40		
General, Non-specific praise				-.27
Extends Discourse		.37		
No Homework/Feedback		.38		
Circulates Inadequately		.28		
Gives Def/Example Only		.38		
Cause or Effect Only		.29		
Does Not State/Apply Rule		.24		
Value Judgement (No Criteria)		.27		
Vague,Scrambled Discourse		.37		
Grating,Inaudible Talk		.26		
Deadpan,Lethargic		.29		
Delays Desist/Punitive		.44		
Loses Momentum		.50		

TABLE 9

Factor Structure
Zero-Order Correlations
FPMS Summative Instrument
Post-Elementary Level

	Factor			
Item	1	2	3	4
Begins Instruction Promptly				
Handles Material Orderly	.53			
Orients/Maintains Focus	.52			
Beginning/Ending Review	.46		.32	
Single Factual Questions				.61
Analysis/Reasoning			.47	.47
Corrective Feedback	.31		.32	.67
Specific Academic Praise	.45		.42	.35
Provide for Practice	.57			
Homework/Seatwork Assign	.53			
Circulates and Assists	.55			
Concept Treatment	.31		.56	
Discuss Cause/Effect			.60	
State/Apply Academic Rule	.31		.39	
Value Judgement			.51	
Emphasize Important Points	.47		.53	
Express Enthusiasm/Challenge	.61	.30	.54	
Body Behavior Shows Interest	.56		.47	
Stops Misconduct				
Maintain Momentum	.44			
Delays		.41		
Not Organized		.44		
Talk Unrelated to Topic		.51		
Multiple Questions,Unison Resp.		.29		-.28
Non-academic Questions		.32		
Ignoring,Sarcasm,Disgust		.43		
General, Non-specific praise				-.26
Extends Discourse		.38		
No Homework/Feedback		.42		
Circulates Inadequately		.31		
Gives Def/Example Only		.36		
Cause or Effect Only		.27		
Does Not State/Apply Rule				
Value Judgement (No Criteria)				
Vague,Scrambled Discourse		.39		
Grating,Inaudible Talk		.27		
Deadpan,Lethargic		.32		
Delays Desist/Punitive		.51	.27	
Loses Momentum		.56		

TABLE 10

Final Estimates of Explained Variance
FPMS Summative Instrument
Post-Elementary Level

Factor	Variance Explained	
	Ignoring Other Factors	Eliminating Other Factors
1	3.67	2.00
2	3.35	2.38
3	3.15	1.40
4	1.91	1.22

TABLE 11

Eigenvalues of Four Factors Retained
FPMS Summative Instrument
Elementary Level

Factor	Eigenvalue	Proportion of Variance	Cumulative Proportion
1	4.811	0.57	0.57
2	2.141	0.25	0.82
3	1.194	0.14	0.96
4	0.810	0.09	1.05

TABLE 12

Factor Correlations
FPMS Summative Instrument
Elementary Level

	Factor 2	Factor 3	Factor 4
Factor 1	.30	.45	.37
Factor 2		.36	.07
Factor 3			.10

TABLE 13

Reference Structure
Semi-partial Correlations
FPMS Summative Instrument
Elementary Level

	Factor			
Item	1	2	3	4
Begins Instruction Promptly				
Handles Material Orderly	.46			
Orients/Maintains Focus	.42			
Beginning/Ending Review	.24			
Single Factual Questions				.55
Analysis/Reasoning			.30	.34
Corrective Feedback				.57
Specific Academic Praise	.28			
Provide for Practice	.47			
Homework/Seatwork Assign	.40			
Circulates and Assists	.44			
Concept Treatment			.37	
Discuss Cause/Effect			.59	
State/Apply Academic Rule			.35	
Value Judgement			.53	
Emphasize Important Points	.35		.25	
Express Enthusiasm/Challenge	.45			
Body Behavior Shows Interest	.45			
Stops Misconduct	.26			
Maintain Momentum	.42			
Delays		.29		
Not Organized		.41		
Talk Unrelated to Topic		.45		
Multiple Questions,Unison Resp.		.24		-.28
Non-academic Questions		.25		
Ignoring,Sarcasm,Disgust		.36		
General, Non-specific praise				
Extends Discourse		.39		
No Homework/Feedback		.44		
Circulates Inadequately		.31		
Gives Def/Example Only		.35		
Cause or Effect Only		.25		
Does Not State/Apply Rule		.30		
Value Judgement (No Criteria)		.21		
Vague,Scrambled Discourse		.37		
Grating,Inaudible Talk		.28		
Deadpan,Lethargic		.32		
Delays Desist/Punitive		.43		
Loses Momentum		.53		

TABLE 14

Factor Structure
Zero-Order Correlations
FPMS Summative Instrument
Elementary Level

	Factor			
Item	1	2	3	4
Begins Instruction Promptly				
Handles Material Orderly	.54			
Orients/Maintains Focus	.48			
Beginning/Ending Review	.43		.34	
Single Factual Questions				.60
Analysis/Reasoning			.36	.38
Corrective Feedback	.35			.65
Specific Academic Praise	.49		.35	.31
Provide for Practice	.54			
Homework/Seatwork Assign	.50		.31	
Circulates and Assists	.46			
Concept Treatment	.39		.50	
Discuss Cause/Effect			.63	
State/Apply Academic Rule	.32		.46	
Value Judgement			.56	
Emphasize Important Points	.58		.49	
Express Enthusiasm/Challenge	.64	.29	.40	
Body Behavior Shows Interest	.61		.36	
Stops Misconduct				
Maintain Momentum	.49			
Delays		.32		
Not Organized		.46		
Talk Unrelated to Topic		.50		
Multiple Questions,Unison Resp.		.32		
Non-academic Questions		.33		
Ignoring,Sarcasm,Disgust		.40		
General, Non-specific praise				.27
Extends Discourse		.40		
No Homework/Feedback		.47		
Circulates Inadequately		.33		
Gives Def/Example Only		.37		
Cause or Effect Only				
Does Not State/Apply Rule		.31		
Value Judgement (No Criteria)				
Vague,Scrambled Discourse		.39		
Grating,Inaudible Talk		.29		
Deadpan,Lethargic		.34		
Delays Desist/Punitive		.49		
Loses Momentum		.59		

TABLE 15

Final Estimates of Explained Variance
FPMS Summative Instrument
Elementary Level

Factor	Variance Explained	
	Ignoring Other Factors	Eliminating Other Factors
1	3.96	1.92
2	3.55	2.39
3	2.95	1.33
4	1.72	1.03

TADS Instrument

Tables 16 and 17 present, respectively, the eigenvalues and inter-factor correlations for the TADS instrument analysis. Factor loadings are presented in Table 18 and item-factor correlations follow in Table 19.

The first factor on the TADS instrument appears to be loading items indicative of content substance and organization, verbal interaction and teacher affect. Highest loadings are on question and comment seeking, the elaboration of student ideas, providing a motivating introduction to the lesson, "personalized" lessons, and providing personal recognition. Although the match is not perfect, parallels may be drawn between this factor and Factor one of the FPMS. As seen in the FPMS analysis, items measuring teacher affect load together teacher efficiency and organization items.

Factor two is a behavior control factor. Activities which maintain or re-direct student attention to lessons, estimated task engagement, and classroom behavior management items load on this factor.

Factor three contains a variety of items with low magnitudes of factor loadings. The items with the most substantial loadings (greater than 0.30) are affective items: patience, fairness, courtesy, and voice tone.

The fourth and final factor retained in this analysis is one comprising the selection and use of materials and teaching aids for the observed lesson. Items loading on this factor concern the appropriateness of materials for both learners and lesson objectives, the teacher's skill in the use of the materials, and the perceived effectiveness of the materials during the lesson (enhancing accomplishment of the objectives).

Table 20 presents final estimates of explained variance for each factor of the TADS instrument. Factor one is the most important, followed by Factor two, Factor four, and Factor three. The large differences between absolute estimates and estimates controlling for the other factors reflects the high degree of inter-factor correlations (Table 15) -- most variance may not be unambiguously attributed to a single factor, but is shared among the factors.

TABLE 16

Eigenvalues of Four Factors Retained
TADS Instrument

Factor	Eigenvalue	Proportion of Variance	Cumulative Proportion
1	16.374	0.59	0.59
2	2.799	0.10	0.69
3	2.107	0.08	0.77
4	1.538	0.05	0.82

TABLE 17
Factor Correlations
TADS Instrument

	Factor 2	Factor 3	Factor 4
Factor 1	.51	.53	.45
Factor 2		.47	.36
Factor 3			.40

TABLE 18

Reference Structure
Semi-partial Correlations
TADS Instrument

	Factors			
Item	1	2	3	4
No lack of knowledge errors			.20	
Information up-to-date			.19	
Important dimensions of topics	.31			
More than one cognitive level	.29			
Logical sequence of information			.29	.20
Instruction appropriate to needs			.21	
Opportunity to practice				
More than passive listening	.20			
More than one group size	.24			
Personalized lesson	.42			
Aids are appropriate for learners				.64
Aids are appropriate for obj.				.65
Aids used at right time				.62
Aids used skillfully				.49
Aids enhance objectives	.21			.49
Materials appropriate for learners			.24	.32
Materials appropriate for obj.			.26	.34
Materials used at right time			.25	.33
Supplemental materials used	.34			.22
Materials enhance objectives	.28			.31
Motivating introduction	.39			
Background is established				
Logical sequence of components			.29	
Appropriate closure of lesson	.32			
Learner attention ensured		.42		
Clear explaination of content			.25	
Precise communication			.26	
Major points emphasized	.32			
Examples/demonstrations used	.27			
No confusion evident				
Clarification			.27	
Words/ideas in clarification	.24			
Individual clarification				
Effective clarification				
Class contribution acknowledged			.23	
Comments/questions sought	.42			
Response sufficiency	.22		.21	
Elaborate learner ideas	.48			
Monitor class performance	.22			
Solicit response for assessment	.31			
Levels of learning monitored	.36			

TABLE 18 (Continued)

Reference Structure
Semi-partial Correlations
TADS Instrument

	Factors			
Item	1	2	3	4
Class evaluates own performance	.35			
Probing	.39			
Expectations communicated	.25			
Specific feedback (+)	.34			
Specific feedback (-)	.38			
Suggestions for improvement	.40			
Begins promptly				
No unnecessary delays		.28		
No undesirable digressions		.27		
Activities fit time period				
Attention ensured		.46		
Clear directions		.26	.22	
Materials on hand and ready			.20	
Efficient handling of routine		.23	.21	
85% or more students on task		.41		
Stimuli are varied	.27			
Active involvement sought	.34			
Non-verbal techniques		.44		
Verbal techniques		.41		
Successful efforts		.59		
Maintain attention		.58		
Behavior expectations		.45		
Consistent expectations		.50		
Behavior is monitored		.47		
Verbal/Non-verbal feedback	.24	.22		
Interference dealt with quickly		.61		
Interference dealt with appropr.		.61		
Equal opportunity to participate			.29	
Learners are encouraged	.34			
Involvement sought	.40			
Personal recognition	.40			
Fairness			.41	
Patience,empathy,understanding			.37	
No sarcasm/ridicule			.31	
Climate of courtesy and respect			.33	
Enthusiasm	.38			
Importance of content stated	.39			
Positive tone of voice			.40	
Knows and uses student names			.28	
Smiling,laughing	.28		.25	
Sitting,standing near students			.22	

TABLE 19

Factor Structure
Zero-Order Correlations
TADS Instrument

Item	Factors 1	2	3	4
No lack of knowledge errors			.32	
Information up-to-date			.35	.29
Important dimensions of topics	.49			.36
More than one cognitive level	.48		.33	.33
Logical sequence of information			.47	.39
Instruction appropriate to needs	.37		.43	
Opportunity to practice	.38		.38	
More than passive listening	.39			.34
More than one group size	.35			
Personalized lesson	.51			
Aids are appropriate for learners				.69
Aids are appropriate for obj.				.73
Aids used at right time				.72
Aids used skillfully				.63
Aids enhance objectives				.65
Materials appropriate for learners			.42	.46
Materials appropriate for obj.			.43	.48
Materials used at right time			.45	.49
Supplemental materials used	.48			.39
Materials enhance objectives	.51			.50
Motivating introduction	.50			
Background is established	.34		.36	
Logical sequence of components			.48	.38
Appropriate closure of lesson	.45			
Learner attention ensured		.56		
Clear explaination of content	.36	.36	.45	
Precise communication			.40	
Major points emphasized	.52		.36	
Examples/demonstrations used	.44			
No confusion evident	.45	.41	.42	
Clarification			.45	
Words/ideas in clarification	.46		.39	
Individual clarification	.38		.38	
Effective clarification	.46	.43	.44	
Class contribution acknowledged			.37	
Comments/questions sought	.54			
Response sufficiency	.47		.45	
Elaborate learner ideas	.58			
Monitor class performance	.39		.33	
Solicit response for assessment	.47		.33	
Levels of learning monitored	.55			

TABLE 19 (Continued)

Factor Structure
TADS Instrument

Item	Factors 1	2	3	4
Class evaluates own performance	.44			
Probing	.58	.37	.36	
Expectations communicated	.41			
Specific feedback (+)	.48		.32	
Specific feedback (-)	.53		.38	
Suggestions for improvement	.53	.30	.31	
Begins promptly				
No unnecessary delays		.43	.36	
No undesirable digressions		.44	.40	
Activities fit time period	.30	.33	.32	
Attention ensured	.31	.58	.31	
Clear directions	.31	.45	.44	
Materials on hand and ready			.33	
Efficient handling of routine		.39	.39	
85% or more students on task	.35	.57	.34	
Stimuli are varied	.46	.33	.34	
Active involvement sought	.53	.40	.34	
Non-verbal techniques	.41	.59		
Verbal techniques	.35	.55	.30	
Successful efforts	.41	.72	.30	
Maintain attention	.41	.70		
Behavior expectations		.56	.34	
Consistent expectations	.34	.63	.35	
Behavior is monitored	.38	.62	.31	
Verbal/Non-verbal feedback	.43	.41	.30	
Interference dealt with quickly		.67		
Interference dealt with appropr.		.68		
Equal opportunity to participate			.39	
Learners are encouraged	.50		.41	
Involvement sought	.55	.32	.34	
Personal recognition	.50		.33	
Fairness			.47	
Patience,empathy,understanding	.36		.51	
No sarcasm/ridicule			.39	
Climate of courtesy and respect		.35	.47	
Enthusiasm	.58	.33	.39	.31
Importance of content stated	.49			
Positive tone of voice	.37		.52	
Knows and uses student names	.30		.39	
Smiling,laughing	.42		.41	
Sitting,standing near students	.35		.36	

TABLE 20

Final Estimates of Explained Variance
TADS Instrument

Factor	Variance Explained	
	Ignoring Other Factors	Eliminating Other Factors
1	13.04	4.06
2	10.71	3.62
3	10.25	2.34
4	7.91	2.54

Number of Observers Used in the 1984-1985 MTP

One of the concerns in starting the Master Teacher Program (MTP) was the availability of observers. Board rule specified that observers had to be trained, tested, and approved prior to conducting observations for the MTP. This ruled out the standard practice of simply appointing persons to conduct observations for purposes of teacher evaluation. Preparations, therefore, had to be made to provide an adequate number of approved observers to implement the MTP.

Four different instruments -- FPMS, TADS, Library/Media and Student Services -- were used in the performance evaluations of candidates in the MTP which required preparation of observers for each instrument.

Records in the Teacher Evaluation and Assessment Center (TEAC) show that by August, 1985 the available pool of observers for each instrument was as follows:

Instrument	Number of Observers
1. FPMS	9,211
2. TADS	809
3. Library/Media	901
4. Student Services	1,184

The number of observers actually used in the MTP is fewer than the number of approved observers available (Table 21). We conclude from this that an adequate number of observers were available to service the MTP in its first year of operation. This prevailed even though over 24,000 candidates were observed from one to four times during this period.

In contemplating resource requirements for future years in the MTP, or similar programs, we conducted analyses to determine the average number of observations each observer conducted. The mean number ranged from 1.7 for Library/Media instrument to 6.4 for the FPMS instrument. The minimum number of observations conducted was 1.0 which was also the mode for all four instruments. The maximum number was in the 50's and 60's on the FPMS and TADS instruments (Table 21).

The number of observers used in the MTP in its first year of operation was also determined by county. Table 22 shows these numbers as the total number of observers used for all four instruments in each county. Table 23 shows the number by instrument in each county. It can be seen that two counties, Broward and Dade, engaged over 1000 observers in the MTP in 1984-1985, while several counties including Flagler, Gilcrest, Glades, etc. engaged under ten.

Were more observers prepared than needed in the MTP? The data cited above show more approved observers than those used in the program. However, observers prepared to use these instruments also conducted teacher evaluations for the Beginning Teacher Program, and, in some cases, these instruments were used in school districts for purposes of annual evaluation. We further assume that learning to conduct observations, which also entails learning the research base for the instrument, has an in-service training effect on the observers. While learning to observe, they are studying the literature on effective teaching and while observing, they are learning the techniques used by other teachers. These experiences may have an impact on the observers' ability to teach and/or to more objectively confer with teachers on their observed performance.

TABLE 21

Observers Participating in 1984/1985 Master Teacher Program

	Instruments				
	FPMS	TADS	MEDIA	ST. SERVICES	TOTAL
N of Observers	6444	1003	433	1079	8132
N of Observations Per Observer					
Mode	1.0	1.0	1.0	1.0	1.0
Mean	6.4	7.4	1.7	3.4	6.5
Median	4.0	5.0	1.0	2.0	4.0
Maximum	61.0	56.0	16.0	25.0	61.0
Minimum	1.0	1.0	1.0	1.0	1.0

TABLE 22

Number of Observers Per County
All Instruments

County	Number of Observers
Alachua	130
Baker	14
Bay	109
Bradford	17
Brevard	264
Broward	1044
Calhoun	11
Charlotte	37
Citrus	61
Clay	88
Collier	81
Columbia	56
Dade	1046
DeSoto	14
Dixie	11
Duval	525
Escambia	181
Flagler	8
Franklin	18
Gadsden	46
Gilchrist	7
Glades	3
Gulf	9
Hamilton	11
Hardee	29
Hendree	21
Hernando	27
Highlands	40
Hillsborough	502
Holmes	14
Indian River	52
Jackson	67
Jefferson	6
Lafayette	4
Lake	84
Lee	180
Leon	118
Levy	22
Liberty	8
Madison	18
Manatee	108
Marion	109
Martin	56
Monroe	38
Nassau	50
Okaloosa	179
Okeechobee	28
Orange	406
Osceola	41

TABLE 22 (Continued)

Number Of Observers Per County
All Instruments

County	Number of Observers
Palm Beach	417
Pasco	137
Pinellas	298
Polk	322
Putnam	45
St. Johns	63
St. Lucie	66
Santa Rosa	76
Sarasota	125
Seminole	207
Sumter	25
Suwannee	21
Taylor	26
Union	7
Volusia	162
Wakulla	11
Walton	37
Washington	25
Fl School Deaf & Blind	19
PK Yonge Lab School	6
FL State Dev Rsch School	9
AD Henderson Univ School	13
FL A&M Univ School	2

TABLE 23

Number Of Observers By Instrument Per County
FPMS Instruments

County	Number of Observers
Alachua	126
Baker	14
Bay	100
Bradford	14
Brevard	236
Broward	890
Calhoun	11
Charlotte	37
Citrus	59
Clay	86
Collier	79
Columbia	54
DeSoto	12
Dixie	11
Duval	465
Escambia	159
Flagler	8
Franklin	14
Gadsden	40
Gilchrist	7
Glades	3
Gulf	7
Hamilton	11
Hardee	26
Hendree	20
Hernando	34
Highlands	39
Hillsborough	465
Holmes	11
Indian River	46
Jackson	64
Jefferson	6
Lafayette	4
Lake	82
Lee	150
Leon	105
Levy	21
Liberty	8
Madison	16
Manatee	101
Marion	99
Martin	54
Monroe	35
Nassau	46
Okaloosa	150
Okeechobee	27
Orange	393
Osceola	39
Palm Beach	391
Pasco	118

TABLE 23 (Continued)

Number Of Observers By Instrument Per County
FPMS Instruments

County	Number of Observers
Pinellas	267
Polk	308
Putnam	45
St. Johns	61
St. Lucie	58
Santa Rosa	67
Sarasota	117
Seminole	194
Sumter	23
Suwannee	17
Taylor	25
Union	6
Volusia	149
Wakulla	11
Walton	35
Washington	21
Fl School Deaf & Blind	16
PK Yonge Lab School	6
FL State Dev Rsch School	8
AD Henderson Univ School	10
FL A&M Univ School	7

TABLE 23 (Continued)

Number Of Observers By Instrument Per County
Media Instruments

County	Number of Observers
Alachua	14
Bay	7
Bradford	2
Brevard	16
Broward	47
Citrus	2
Clay	7
Collier	2
Columbia	3
Dade	62
DeSoto	2
Duval	22
Escambia	8
Flagler	2
Franklin	2
Gadsden	2
Hernando	3
Highlands	4
Hillsborough	6
Holmes	2
Indian River	4
Jackson	2
Lake	3
Lee	12
Leon	5
Manatee	5
Marion	2
Martin	4
Monroe	2
Nassau	3
Okaloosa	15
Okeechobee	4
Orange	23
Osceola	2
Palm Beach	13
Pasco	13
Pinellas	11
Polk	14
Putnam	2
St. Johns	2
St. Lucie	11
Santa Rosa	3
Sarasota	4
Seminole	25
Sumter	2
Suwannee	4
Taylor	2
Volusia	17

TABLE 23 (Continued)

Number Of Observers By Instrument Per County
Media Instruments

County	Number of Observers
Washington	2
FL State Dev Rsch School	2
AD Henderson Univ School	3
FL A&M Univ School	2

Number Of Observers By Instrument Per County
Student Services Instruments

County	Number of Observers
Alachua	16
Baker	2
Bay	9
Bradford	3
Brevard	72
Broward	171
Charlotte	2
Citrus	7
Clay	13
Collier	4
Columbia	6
Dade	182
Duval	42
Escambia	44
Flagler	2
Franklin	3
Gadsden	4
Gilchrist	3
Gulf	2
Hardee	3
Hendry	4
Highlands	5
Hillsborough	39
Holmes	2
Indian River	5
Jackson	3
Lake	6
Lee	25
Leon	12
Levy	6
Liberty	2
Madison	2
Manatee	16

TABLE 23 (Continued)

Number Of Observers By Instrument Per County
Student Services Instruments (Continued)

County	Number of Observers
Marion	16
Martin	16
Monroe	5
Nassau	12
Okaloosa	22
Okeechobee	5
Orange	53
Osceola	6
Palm Beach	26
Pasco	17
Pinellas	25
Polk	29
Putnam	14
St. Johns	3
St. Lucie	22
Santa Rosa	9
Sarasota	8
Seminole	33
Sumter	2
Suwannee	2
Taylor	2
Union	2
Volusia	24
Wakulla	2
Walton	2
Washington	2
FL School Deaf & Blind	3

Number of Observers By Instrument Per County
TADS Instruments

County	Number of Observers
Dade	1003

CHAPTER 6

Special Instruments Norming Studies

With the passage of the Master Teacher Program legislation in 1984, it was recognized that the Florida Performance Measurement System (FPMS) summative instrument, based on effective classroom practice, was inadequate for observing and evaluating student services personnel (occupational placement specialists, guidance counselors, school psychologists, social workers and student services administrators), and library/media specialists. Although classified as instructional personnel and, therefore, eligible to participate in the Master Teacher Program; these specialists do not work in typical instructional settings, but rather student services personnel work in consultative and diagnostic settings while library/media specialists have a number of non-instructional functions not assessed by the FPMS summative instrument.

Following the same procedures used in the development of the FPMS materials, task forces of specialists in each of these areas were formed to develop specialized domains and assessment instruments for these areas. The Student Services Task Force compiled a set of concepts and indicators from a search of counseling and consultation literature. Content validity was established through literature search, judgment of 464 specialists in the field and the consensus of a review team. An observation instrument with 19 effective and 18 ineffective indicators was developed and submitted to studies of reliability. The results of the reliability studies showed that total scores should be used for this instrument. For total score, intraclass discriminant reliability was r = .58; coefficient alpha (internal consistency), r = .78; and intercoder agreement, r = .56.

'Using both professional consensus and a review of the literature, the Library/Media specialists identified nine concepts and thirty-two indicators, with examples and supportive evidence. A cross-section of 727 library/media specialists representing all levels, sizes, and organizations of schools, and all levels of experience, responded to a content validation study with a very high level of agreement on the items included. An instrument was developed from these concepts and indicators that was submitted to tests of reliability. Reliability studies again indicated that the total instument score was the most reliable (r = .77 to .89).

These instruments were submitted to analyses of frame factor effects on scores for purposes of norming for the Master Teacher Program. The findings of this study apply only to the

Master Teacher Program since candidates were self selected based upon program criteria for eligibility and the instrument scores were used to rank order candidates on a scale of effectiveness for purposes of an award program.

This section of the report contains analyses of the relationship between total scores and selected frame factors on Student Services and Library/Media observations conducted in the 1984/85 Master Teacher Program (MTP). These analyses were conducted to identify significant score variations associated with frame factors measured with these instruments. The appearance of significant score variations would indicate the need for creating norm groups or score adjustments to provide equitable assessment of applicant performance independent of the context of evaluation.

Sample

The data comprising the sample consisted of all 1984-1985 MTP observations received prior to August 1, 1985. The following are the total samples available by instrument:

Single Observations

Student Services	3,577
Library/Media	717

Analytic Design

For the analyses of frame factors on the Student Services and Library/Media instruments, the observations were randomly divided into two groups using the SAS UNIFORM(0) function to allow cross-validation of results, and to provide two tests of significance using unique data for each frame factor. Continuous variables (e.g. applicant's number of years of experience) were grouped by quartiles and the differences among the four group means were analyzed using one-way analyses of variance.

A set of criteria identical to that used in the norming of the FPMS summative instrument was used to interpret these analyses.

Specifically:

1. Differences among group means were at least equal in magnitude to one-third standard deviation for the instrument.
2. An alpha level of 0.01 was set for each test, to somewhat reduce the experiment error rate inflated by the large number of statistical tests.

3. Significant group differences were identified in each random sample.
4. A minimum cell size of 100 was present to avoid generalizations based upon few cases.

Student Services Frame Factor Study

Eight frame factors are listed on the Student Services Instrument. For clarity of presentation in this report, the frame factors were organized into three groups:

1) Personal attributes of applicants
   - Sex
   - Racial/ethnic group
2) Professional attributes of applicants
   - Instructional classification
   - Years of experience
3) Characteristics of the setting
   - Type of session
   - Nature of participants
   - Level of students
   - Number of participants

Table 1 presents the analysis of personal attributes of applicants. Female applicants received higher mean scores than male applicants in both samples ($p<0.01$). The mean differences, however, were less than the 1/3 standard deviation criterion (the differences between means being 1.0 and 0.7 for Samples 1 and 2 respectively). On the factor of racial/ethnic group, black applicants received significantly lower scores on Sample 2, but this result was not present in Sample 1 (Table 1).

TABLE 1

Student Services Instrument
Score Variation by Frame Factors
Personal Attributes of Applicants

	Sample 1			Sample 2		
	N	Mean	S.D.	N	Mean	S.D.
Sex						
Male	550	25.5	5.2	578	25.7	5.2
Female	1175	26.5	4.9	1235	26.4	4.8
	F=15.84 (p<0.01)			F=9.99 (p<0.01)		
Race/Ethnic Group						
White	1442	26.3	4.9	1500	26.4	4.8
Black	177	25.6	5.5	222	24.9	5.7
Hispanic	80	26.0	5.6	83	26.7	4.7
Other	10	25.7	4.4	3	28.3	4.2
	F=0.81 (p<0.49)			F=6.44 (p<0.01)		

Table 2 presents the analysis of applicants' professional attributes. Most applicants were classified as counselors (71% of the total), although all five categories were well represented. In neither subsample were mean differences significant on this variable. However, given the substantial number of candidates coding the "Other" category of instructional classification, this frame factor should be examined to determine if additional categories of instructional classification should be added (Table 2).

Similarly, years of experience appears unrelated to instrument scores, but it is interesting to note that in both subsamples, the highest experience group received the lowest mean score (Table 2).

TABLE 2

Student Services Instrument
Score Variation by Frame Factors
Professional Attributes of Applicants

	Sample 1			Sample 2		
	N	mean	S.D.	N	mean	S.D.
Instructional Classification						
Counselor	1223	26.3	5.0	1301	26.3	5.0
Occ. Specialist	61	25.1	5.0	60	25.4	5.6
Psychologist	211	25.5	4.7	228	25.9	4.6
Social Worker	68	25.3	5.0	73	25.5	4.7
Other	179	26.4	5.1	168	26.7	4.4
	F=2.61 (p<0.03)			F=1.62 (p<0.17)		
Applicant's Years of Experience						
0-4	348	26.3	4.9	375	26.4	4.8
5-9	519	26.4	4.9	528	26.3	4.9
10-13	431	26.2	4.9	475	26.3	4.9
>13	432	25.8	5.3	442	25.8	5.2
	F=1.33 (p<0.26)			F=1.08 (p<0.36)		

The analysis of the four characteristics of the settings in which the observations were conducted is presented in Table 3. Three categories of session type were recorded: Counseling, consultation, and other. The "Other" category is characterized by a significantly lower mean score than the remaining categories, although the magnitude of the differences among means falls short of the criterion. Again, because of a large proportion of "Other" session types, this frame factor should be examined to determine if additional categories should be specified.

The nature of the participants (Students, faculty, parents, other) appears unrelated to instrument scores, as does the grade level of students.

In both subsamples, the number of participants was significantly related to scores on the Student Services Instrument, with large groups (n>5 participants) receiving the lowest scores. However, the magnitude of the group differences falls short of the one-third standard deviation criterion.

TABLE 3

Student Services Instrument
Score Variation by Frame Factors
Characteristics of the Setting

	Sample 1			Sample 2		
	N	Mean	S.D.	N	Mean	S.D.
Type of Session						
Counseling	1028	26.6	5.1	1099	26.5	5.1
Consultation	504	25.6	4.8	517	26.2	4.6
Other	209	25.3	5.1	206	24.9	4.9
	F=10.41 (p<0.01)			F=8.94 (p<0.01)		
Nature of Participants						
Students	1106	26.4	5.1	1199	26.2	5.1
Faculty	281	26.0	4.7	257	26.3	4.7
Parents	190	24.8	4.6	217	25.7	4.3
Other	166	26.1	5.3	157	26.5	4.9
	F=5.89 (p<0.01)			F=0.90 (p<0.44)		

TABLE 3 (continued)

Student Services Instrument
Score Variation by Frame Factors
Characteristics of the Setting

	Sample 1			Sample 2		
	N	Mean	S.D.	N	Mean	S.D.
Level of Students						
Elementary	660	26.0	4.7	675	26.0	4.7
Junior High	411	26.7	5.2	422	26.5	4.7
Senior High	542	26.0	5.1	594	26.2	5.2
Adult	107	25.2	5.3	107	26.2	6.2
Multi-level	17	26.5	4.5	17	26.5	3.2
	F=2.61 (p<0.03)			F=0.59 (p<0.67)		
Number of Participants						
1	491	26.1	5.1	529	26.4	5.0
2	512	27.0	4.9	565	27.0	4.8
3-5	379	26.1	5.0	377	25.8	4.7
>5	360	25.2	4.9	358	25.0	5.0
	F=8.48 (p<0.01)			F=13.90 (p<0.01)		

Summary of Student Services Analysis

None of the frame factors provide justification for additional norm groups or score adjustments on the Student Services Instrument. A few variables (e.g. sex, race and instructional classification) should be reviewed as additional data become available since these variables suggest a possible relationship with scores. However, none of the variables met all of the criteria for creating additional norm groups or score adjustments in this study.

Library/Media Frame Factor Study

Eight frame factors are noted on the Library/Media Instrument. In a manner similar to that used for the Student Services Instrument, the frame factors are organized into three groups:

1) Personal attributes of applicants
   - Sex
   - Racial/ethnic group
2) Professional attributes of applicants
   - Primary responsibility
   - Years of experience
3) Characteristics of the working environment
   - School level
   - Number of professional personnel in the center
   - Number of support staff in the center
   - Number of students in the school

Table 4 presents descriptive statistics and results of the analyses of variance for the personal attribute variables. Two personal attributes of applicants were recorded on the Library/Media Instrument: sex and racial/ethnic group. Although females scored slightly higher than males and showed less variability in scores, there was no significant difference among means in either subsample. Similarly, white applicants typically scored somewhat higher than other racial/ethnic categories but differences were non-significant (Table 4).

TABLE 4

Library/Media Instrument
Score Variation by Frame Factors
Personal Attributes of Applicants

	Sample 1			Sample 2		
	N	Mean	S.D.	N	Mean	S.D.
Sex						
Male	22	18.9	5.2	25	18.1	6.5
Female	322	20.6	4.8	347	19.7	5.7
	F=2.65 (p<0.10)			F=1.79 (p<0.18)		
Race/Ethnic Group						
White	330	19.6	5.7	309	20.7	4.8
Black	35	19.3	5.2	30	18.7	4.6
Hispanic	6	16.5	7.8	4	18.3	7.0
Other	1	25.0	---	1	25.0	---
	F=0.91 (p<0.44)			F=2.20 (p<0.09)		

The two most common "primary responsibilities" of the Library/Media applicants for the MTP were *Print & AV and Print, AV, and ITV*, which together comprised 82% of the total (Table 5). Four categories of experience were created based upon the quartile points of this distribution. Although one of the subsamples yielded a statistically significant group difference on the experience variable, the effect was not present in the second sample. On both of the samples, however, the group with the greatest number of years of experience received somewhat lower mean scores than less experienced groups.

TABLE 5

Library/Media Instrument
Score Variation by Frame Factors
Professional Attributes of Applicants

	Sample 1			Sample 2		
	N	Mean	S.D.	N	Mean	S.D.
Applicant's Primary Responsibility						
Print	32	20.3	4.1	28	20.8	4.1
AV	17	16.1	6.2	20	20.2	4.4
ITV	2	19.5	3.5	1	25.1	---
Print + AV	151	19.4	5.8	104	20.6	4.7
Print + ITV	0	----	---	0	----	---
AV + ITV	8	20.8	6.7	12	20.3	4.0
All	156	19.9	5.8	173	20.5	5.2
Other	5	18.6	4.4	4	18.8	4.0
	F=1.31 (p<0.25)			F=0.27 (p<0.95)		
Applicant's Years of Experience						
0-6	97	19.6	5.7	77	20.3	4.3
7-11	98	20.1	5.6	95	21.7	3.8
12-15	73	20.0	5.4	96	20.8	4.9
>15	102	18.9	5.7	75	19.0	6.0
	F=0.97 (p<0.41)			F=4.57 (p<0.01)		

The results of the analyses of characteristics of the working environment are presented in Table 6. School level was recorded in four categories (Elementary, Junior High, High School, and Other). Among these categories, "Other" received the highest mean rating in both subsamples, although the overall differences among means were not significant in either sample.

Three additional school characteristics were recorded, each of which is an index of school size: number of professional personnel in the library/media center, number of support staff members, and number of students in the school. The number of professional personnel ranged from 1 to 4, and provided the four categories for the ANOVA. The number of support staff was trichotomized (categories of 0, 1, greater than 1) for analysis. The number of students in the

school was categorized based upon the quartile points of its distribution, as described previously. None of the school size variables were significantly related to scores on the Library/Media Instrument (Table 6).

TABLE 6

Library/Media Instrument
Score Variation by Frame Factors
Characteristics of Working Environment

	Sample 1			Sample 2		
	N	Mean	S.D.	N	Mean	S.D.
School Level						
Elementary	186	19.5	5.8	159	20.1	5.7
Junior High	72	19.2	6.7	72	20.8	3.9
High School	93	19.8	5.1	83	20.6	4.0
Other	21	20.7	4.4	29	22.1	3.9
	F= 0.48 (p<0.70)			F= 1.66 (p<0.17)		
Number of Professional Personnel in Library/Media Center						
1	250	19.4	5.9	222	20.3	5.2
2	99	20.0	5.0	92	20.9	4.1
3	14	18.4	6.9	15	20.2	4.5
4	8	20.2	3.6	8	21.4	2.9
	F= 0.56 (p<0.64)			F= 0.43 (p<0.73)		
Number of Support Staff in Library/Media Center						
0	60	18.7	6.1	55	20.1	5.3
1	227	19.9	5.6	205	20.4	5.4
>1	50	19.2	5.7	53	21.1	3.8
	F= 1.14 (p=0.32)			F= 0.59 (p=0.55)		
Number of Students in School						
<= 583	91	19.6	5.7	88	20.6	5.3
584 to 784	93	18.8	5.7	86	19.8	5.4
785 to 1254	105	20.2	6.0	75	20.9	4.8
>= 1255	83	19.5	5.3	95	20.9	3.7
	F= 1.02 (p=0.38)			F= 1.04 (p=0.38)		

Summary of Library/Media Analysis

None of the frame factors provide justification for additional norm groups or score adjustments on the Library/Media Instrument. As occured in the norming study of the FPMS summative instrument, cell sizes for some categories may be inadequate for proper assessment (e.g. only 47 male applicants and 65 black applicants were available for analysis).

CHAPTER 7

Development of Materials For Trainer Update Workshops

An integral arm of the FPMS is the dissemination of training materials and training trainers and observers. Since 1982, the FPMS materials have been used throughout the state of Florida, with the exception of Dade County, and in addition, in the state of Kentucky. All of the materials were validated in field tests (Norming Study 1983-84, Validity Studies 1984-85, and Predictive Validity Studies 1985-86) using procedures that have become a part of the FPMS Training Program. Training materials and procedures have been developed to provide accuracy and consistency in the delivery of the FPMS workshops by trainers who attended demonstration workshops to receive instruction and practice with the training materials. Trainers are certified by passing examinations on the knowledge base of the domain documents, the coding manual, and the coding of criterion videotapes.

Materials for Summative Instrument Update

Observation reliability is a key factor in a measurement system for it is clear that different trainers must code the performance of a teacher in the same way over time or the results are of dubious value. To maintain reliable trainers, workshops to update and monitor the performance of the FPMS corps of certified trainers were planned for 1985-86. We assisted the Florida Department of Education by working with a selected group of trainers who developed plans and procedures for a series of regional workshops focussed on three areas: summative instrument, formative training, and training in Domain 4.0.

In addition to assisting with the design of the first workshop, the following products were readied for dissemination. The summative instrument coding rules as recorded in the FPMS coding manual, June 1984, were updated and more illustrations were added to clarify terms and points of inconsistency. A new trainer's edition of the manual was written with notes to the trainer.

A quiz on the FPMS summative instrument coding manual was developed and field tested, along with a pilot test of the trainers' coding manual, at a workshop conducted in Tampa. Final revisions were made before the first of the regional trainer/observer update workshops. From an

item analysis of the pilot test of 55 items, the quiz was cut to 35 items and administered, along with criterion videotapes to check accuracy of coding, to all trainers who attended the regional workshops in September, 1985.

Materials for Conferring with Teachers

Last year the knowledge and skills needed by a conferrer, that is, principal, supervisor, peer teacher, or anyone who provides help to a teacher whose need is decided on the basis of data derived from observations of classroom performance, were compiled in a FPMS domain. Conference materials in the FPMS manuals prior to last year were drawn from literature on counseling that was not necessarily research based. The development of the knowledge base, the observation instrument, and the coding manual for Domain 10.0, Conferring with Teachers were described in Teacher Evaluation Project, Report 1984-1985. These materials were available as a part of the formative training workshops presented in early 1986.

Training materials were developed for a conferring component of the regional workshops on formative training for subsequent delivery by the trainers in their districts. The prepared materials included 1) the design for a Domain 10.0 workshop, 2) a series of examples of conferrer's statements for practice with a trainer's guide, 3) an audiotape of conferrer's statements with a transcript and trainer's guide, 4) a list of the Domain 10.0 indicators, 5) a design for role playing practice with a role playing checklist, 6) a viewing form for practice with videotapes, and 7) an outline of indicator distinctions (Appendix C). The use of these materials was modeled in the delivery of instruction in the February workshops.

The materials described in this chapter differ from the materials that are used by the Student Services personnel, although the terminology in Domain 10.0 is related to Domains 7.0 and 8.0. The Student Services domains were developed for a special group of teachers who are specifically trained to work in a therapeutic setting.

Domain 10.0 focuses on the behavior of the person who provides help to the teacher. The observation data on the teacher's classroom performance is the primary concern in the conference. The content of the conference is important for unlike the student service worker less attention is

given to theoretical concepts and to student concerns except as they influence the teacher's behavior. Instead attention is given to how the conferrer works with the teacher on a specific data base.

The conference serves as a pedagogical guide as the problems of planning, instruction, student conduct and student evaluation are explored. The conferrer must understand the consulting procedure and possess the knowledge and skills of interpersonal communication as described in Domain 10.0 as well as the pedagogical knowledge and skill to assist the teacher.

The use of this domain and the accompanying materials is very specific. It is a means of helping the conferrer improve his/her skill.

BIBLIOGRAPHY

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., and Tukey, J. W. (1972). **Robust Estimates of Location Survey and Advances.** Princeton, New Jersey: Princeton University Press.

Baldwin, L., Medley, D. M., and MacDougall, M. A. (1983). A comparison of analysis of covariance to within-class regression in the analysis of non-equivalent groups, **Journal of Experimental Education,** 52:1, 68-75.

Bradley, J.W. (1980). Nonrobustness in **Z, T,** and **F** tests at large sample sizes, **Bulletin of the Psychonomic Society,** 16, 333-336.

Brophy, J.E. and Evertson, C.M. (1976). Context Variables in Research on Training, **Correlates of Effective Teaching.** Austin, Texas: The University of Texas, The Research and Development Center for Teacher Education.

Burstein, L., Linn, R. L., and Capell, F. J. (1978). Analyzing multilevel data in the presence of heterogeneous within-class regressions, **Journal of Educational Statistics,** 3:4, 347-383.

Campbell, D.T. and Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix, **Psychological Bulletin,** 56, 81-105.

Cook, T.D. and Campbell, D.T. (1979). **Quasi-experimentation: Design & Analysis Issues for Field Settings.** Chicago: Rand McNally.

Cronbach, L.J. (1970). **Essentials of Psychological Testing,** (3rd ed.). New York: Harper & Row.

Darlington, R.B. (1968). Multiple regression in psychological research and practice, **Psychological Bulletin,** 69, 161-182.

Effron, B. (1982). **The Jacknife, the Bootstrap, and Other Resampling Plans.** Philadelphia, Pa.: Society for Industrial and Applied Mathematics.

Elashoff, J. D. and Elashoff, R. M. in Kruskal, W.H. and Tanur, J. M. (Ed.) (1978). **International Encyclopedia of Statistics,** Vol. 1, NY: Free Press.

Far West Laboratory for Educational Research and Development (1978). **The Beginning Teacher Evaluation Study -- Supplement Phase III-A and Phase III-A Continuation,** San Fransisco, Ca.

Florida Statewide Assessment Program (1982). **Technical Report: for April 1981, October 1981, April 1982.** Tallahassee, Fl: State of Florida, Department of Education.

Florida Statewide Assessment Program (1980). **Technical Report: for October 1979, April 1980, October 1980.** Tallahassee, Fl: State of Florida, Department of Education.

Gastwirth, J.L. (1971). On the sign test for symmetry, **The Journal of the American Statistical Association,** 66, 821-823.

Glass, G. V. (1982). Meta-Analysis: an approach to the synthesis of research results, **Journal of Research in Science Teaching,** 19:2, 93-112.

Hedges, L. V. and Olkin, I. (1983). Clustering estimates of effect magnitude from independent studies, **Psychological Bulletin**, 93:3, 563-573.

Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments, **Psychological Bulletin**, 92:2, 490-499.

Huber, P. J. (1972). The 1972 Wald lecture on robust statistics: a review, **Annals of Mathematical Statistics**, 43:4, 1041-1067.

Jackson, G. B. (1980). Methods for integrative reviews, **Review of Educational Research**, 50:3, 438-460.

Kendall, M. G. (1948). **The Advanced Theory of Statistics, Vol. II**, London: Griffin & Company Ltd.

Kerlinger, F.N. (1967). **Foundations of Behavioral Research**. New York: Holt, Rinehart, and Winston Inc.

Kerlinger, F. N. and Pedhazur, E. J. (1973). **Multiple Regression in Behavioral Research**. San Francisco, Ca.: Holt, Rinehart, and Winston Inc.

Klitgaard, R. E. (1978). Identifying exceptional performance. **Policy Analysis**, 4:4, 529-547.

Klitgaard, R. E., Dadabhoy, S., and Litkouhi, S. (1981). Regression without a model, **Policy Sciences**, 13, 99-113.

Kromrey, J., Micceri, T., Peterson, D., and Smith, B.O. (1986). **Guidelines for Conducting Predictive Validity Studies of Measured Teacher Performance**. unpublished technical report, Tampa: University of South Florida, College of Education.

Light, R. J. and Pillemer, D. B. (1984). **Summing up: The Science of Reviewing Research**. Cambridge, Ma.: Harvard University Press.

Lord, F. M. (1958). Further problems in the measurement of growth, **Educational and Psychological Measurement**.

Micceri, T. (1986). **An Exploratory Comparison of Ex Post Facto Methods for Equating Non-equivalent Groups**. unpublished technical report, Tampa, Fl: University of South Florida, College of Education.

Micceri, T. (1986). **FPMS Validity Studies C and D: Long Term Studies of Elementary Mathematics and Post-elementary Literature**. unpublished technical report, Tampa, Fl: University of South Florida, College of Education.

Micceri, T. (1985). **FPMS Validity Report #1 Overview of Florida Performance Measurement System Predictive Validity Studies A and B: Theoretical Background, Results and Implications**. unpublished technical report, Tampa, Fl: University of South Florida, College of Education.

Micceri, T. (1984a). **A Technical Report on the Florida Performance Measurement System Norming Study**. unpublished technical report, Tampa, Fl: University of South Florida, College of Education.

Micceri, T. (1984b). **Preliminary Report on Beginning Teacher Activities 1983/84**. unpublished technical report, Tampa, Fl: University of South Florida, College of Education.

Micceri, T. (1983). **A Report on the Reliability Study for the Florida Performance Measurement System.** unpublished technical report, Tampa, Fl: University of South Florida, College of Education.

Medley, D.M., Coker, H., and Soar, R.S. (1984). **Measurement-Based Evaluation of Teacher Performance: An Empirical Approach.** New York: Longman, Inc.

Mosteller, F. and Tukey, J.W. (1977). **Data Analysis and Regression: A Second Course in Statistics.** Reading, Ma.: Addison-Wesley.

Miller Analogies Test Manual (1970). New York: The Psychological Corporation.

Rogosa, D., Floden, R. E., and Willett, J. B. (1984). **Assessing the Stability of Teacher Behavior: Research Series No. 141.** (monograph), East Lansing, Mi.: Institute for Research on Teaching.

Rosenthal, R. (1983). Assessing the statistical and social importance of the effects of Psychotherapy, **Journal of Consulting and Clinical Psychology**, 51:1, 4-13.

Rosenthal, R. and Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect, **Journal of Educational Psychology**, 74:2, 166-169.

Rosenthal, R. (1978). Combining results of independent studies, **Psychological Bulletin**, 85:1, 185-193.

Soar, R. (1984). Personal conversation.

Sternberg, R. J. (1984). What should intelligence tests test? Implications of a triarchic theory of intelligence for intelligence testing, **Educational Researcher**, 13:1, 5-15.

Sudman, S. (1976). **Applied Sampling.** NY.: Academic Press.

Teacher Education Internship Project (1984). **Final Report.** Tampa, Fl.: University of South Florida, College of Education.

Teacher Evaluation Project: Report 1982-1983 (Document No. SP 027190). Tampa, Fl: University of South Florida, College of Education. (ERIC Document Reproduction Service No. ED 266 121).

Teacher Evaluation Study: Report for 1983-1984 (Document No. SP 027191). Tampa, Fl: University of South Florida, College of Education. (ERIC Document Reproduction Service No. ED 266 122).

Teacher Evaluation Project, 1984-1985 (Document No. SP 027189). Tampa, Fl: University of South Florida, College of Education. (ERIC Document Reproduction Service No. ED 266 120).

Teacher Evaluation and Assessment Center: Report for 1984-85 (Document No. SP 026680). Tampa, Fl: University of South Florida, College of Education. (ERIC Document Reproduction Service No. ED 266 089).

Veldman, D.J. and Sanford, J.P. (1984). The Influence of Class Ability Level on Student Achievement and Classroom Behavior, **American Educational Research Journal**, 21:3, 629-644.

Wainer, H. (1982). Robust statistics: A survey and some prescriptions, in Keren, G. (ed.), **Statistical and Methodological Issues in Psychology and Social Sciences Research.** Hillsdale, N.J.: Lawrence Erlbaum Associates.

Wade, R. K. (1985). What makes a difference in inservice teacher education? A meta-analysis of research, **Educational Leadership**, 48-54.

APPENDIX A

APPENDIX A

Proposed Changes in State Board Rule for the Beginning Teacher Program

The study of portfolios reported in Chapters 1 and 2 lends support to the following recommendations for changes in State Board of Education Rule.

1. To insure the competence of observers of beginning teacher classroom performance, thereby improving the accuracy and consistency of observations, provision should be made in State Board Rule for state approval (certification).
2. The diagnosis of BT performance should include screening observations by approved observers, analysis of the compiled data in relation to state norm scores, and planning the sequence of formative observations and remediation activities to assist the BT develop competence in all areas of teaching. The professional development plan should specify the sequence of observations, activities, and the details for monitoring and assessing the plan.
3. All members of the BT support staff should be trained in diagnosis, prescription, conferring, and all aspects of clinical supervision as well as the FPMS knowledge base.
4. Released time for clinical supervision should be provided for every peer teacher.
5. The professional development plan and the portfolio of each BT should meet state criteria.
6. Develop criteria for evaluation of inservice activities.
7. Conferences with the BT should include a data-base. Each conference should be based upon at least 2 independent observations by approved observers.
8. Formative observations and conferences should begin early in the academic year.
9. To provide information for developing a quantitative and qualitative data base on the State's beginning teachers, provision should be made in State Board Rule for districts to submit screening and summative observations of beginning teachers to the DOE.
10. A sample of portfolios should be reviewed by the DOE for compliance with BTP requirements.

11. To improve the reliability and validity of evaluation scores by collecting a representative sample of beginning teacher classroom performance at the beginning and end of the school year, provision should be made for a minimum of three screening observations by three different observers and a minimum of three summative observations by three different observers.

A Review of Research on Evaluation of the Portfolio Process

What research has been conducted on the documentation of a beginning teacher's performance? Are there analyses of the documentation, or what now is called the portfolio, in the Florida Performance Measurement System?

The first part of Chapter 1 provided data on the study of a sample of portfolios to look at the contents of the documentation in regards to teacher competency and remediation. We turn now to a search of the literature to find sources that may help determine what changes in the documentation might be advisable in the future.

A literature search of the ERIC data base identified a few references that seem worthy of pursuing. These materials and others which describe the documentation of teacher competencies are listed in the following Bibliography on Portfolios and Professional Development.

Bibliography on Portfolios and Professional Development

Arends, Richard and others (March, 1978). Inservice Education and the 6 O'Clock News (Document No. EAQ10387). OSSC Bulletin. 21:7, 33 p. (ERIC Document Reproduction Service No. ED150746).

Several assumptions currently held by many inservice planners and deliverers are presented and challenged. A new model of inservice programs is outlined that pays attention to the characteristics of maturing and growing professional educators and that demands new forms for delivering inservice to individuals and school clients. Inservice delivery models proposed include the linking agent model and the organization development model. The linking agent model consists of field agents who travel from school to school helping clients identify their needs and then

providing assistance as they search for research validated solutions. The organization development model aims training and development efforts toward helping people in schools come together and use the resources they already possess in more effective and satisfying ways. The paper concludes with a list of policy and organizational changes needed within agencies that currently house the major actors in inservice education: local school districts, state departments of education, and institutions of higher education.

Child Development Associate (1975). Competency Standards. (Printed under contract, No. 105-77-1012). US Department of Health, Education, and Welfare: The Child Development Associate Consortium.

This document describes, in question and answer format, the Credential Award System through which the Child Development Associate (CDA) Consortium regulates assessment of child care personnel who apply for the CDA credential. A private, non-profit organization of national professional associations, the CDA Consortium was founded in 1972 with the aim of developing a system for (1) assessing the competence of individuals working with 3- to 5-year olds in early childhood centers and (2) awarding the CDE credential to persons assessed as competent. This system covers personnel in nursery schools, parent-child cooperatives, day care centers, Head Start programs, and other center-based programs. After minimal eligibility requirements have been satisfied, a candidate is evaluated in six areas of competence in working with children: (1) establishment and maintenance of a safe and healthy learning environment, (2) advancement of physical and intellectual competence, (3) support of a positive self-image and individual strength, (4) promotion of harmonious interrelation between children and adults in a group environment, (5) development of good coordination between home and center with regard to child-rearing practices and expectations, and (6) ability to carry out supplementary responsibilities related to children's programs. Members of a Local Assessment Team -- consisting of the candidate, a trainer, a parent-community representative and a Consortium representative -- collect information on the candidate's performance, coordinate their judgements according to the Consortium's competence standards list, and determine whether the candidate should be awarded a credential.

Commission for teacher preparation and licensing: a report of bilingual/crosscultural educational credentialing requirements (1975). (Document No. FL008318) Sacramento, CA: California State Commission for Teacher Preparation and Licensing, 38 p. (ERIC Document Reproduction Service No. ED134007).

This report, designed to provide the Commission for Teacher Preparation and Licensing with documentation concerning the need for credentialing in bilingual/cross-cultural education, is based

on the examination of the performance of minority populations in California having a cultural and linguistic background different than that of the dominant majority. The historical background of the investigation is provided, along with an overview and definitions of bilingual/cross-cultural education. Educational outcomes providing indices of schools' success in educating minority students include school holding power, reading skill, grade repetition, overageness, and participation in extracurricular activities. A major factor in under-achievement is the lack of teachers drawn from the same minority and/or cultural groups as the students, along with a shortage of minority personnel being prepared to teach. Charts provide statistics on all outcomes. The report recommends that the Commission announce the need for credentialing and identify the standards for credentialing, and that the Bilingual/Cross-Cultural Education Committee identify the types of credentials required, investigate the possibility of creating "career ladder" opportunities to make qualified personnel available immediately, and investigate the requirements for re-training existing certificated personnel in bilingual/cross-cultural education. A preliminary definition and qualifications for a specialist in bilingual/cross-cultural education are provided.

Conine, Tali A. and others (May, 1985). The relative importance of supportive data for promotion and tenure reviews. Journal of Allied Health. 14:2, pp.183-190.

Faculty in 21 schools of allied health in the Midwest were surveyed by a questionnaire to identify (1) the relative importance of the variety of documents allied health faculty may submit for promotion and tenure reviews and (2) the congruence between the real and ideal value of the major areas of faculty responsibility.

Eade, Gordon E. (1980). The competency centered/field based early childhood certification project. National Council of States on Inservice Education Occassional Paper Number 2. Syracuse, New York: National Dissemination Center, 4 p.

Collaboration among representatives from panhandle universities and school districts resulted in a certification project adopted by Bay School District and Okaloosa School District in the fall of 1977. The inservice field-based project included nine procedural steps including one on the development of a portfolio by the teacher-candidate. The portfolio, described as a personalized compilation of information to document ten early childhood competencies, was presented to a support team for evaluation and further planning for inservice study. Ultimately, the support team made a recommendation regarding the candidate's satisfactory completion of certification requirements.

Frels, Kelly and Cooper, Timothy (1983). Documentation of employee performance. (Document No. EA016926). Legal Issues in Public School Employment, Chapter 6, 17 p.(ERIC Document Reproduction Service No. ED245377).

Although the main purpose of a school district's evaluation system is to improve employee's performance, the evaluation procedures and the supporting documentation must also serve as evidence in the termination hearings of an unsatisfactory employee. Accordingly, this chapter provides practical advice to school administrators concerning documentation generated in the evaluation process that can be used, if necessary, to support a recommendation for dismissal. Five types of written memoranda are described, with examples of each: (1) memoranda to the administrator's files (used for minor infractions or deviations by the employee); (2) specific incident memoranda (used to record conferences with a teacher concerning more significant events); (3) summary memoranda (used to record conferences with a teacher in which a series of incidents, problems, or deficiencies are discussed); (4) visitation memoranda (used for observing on-the-job performance); (5) formal assessment instruments (used to evaluate overall performance). The latter part of the chapter discusses factors to consider in termination proceedings, including timely notification, opportunity for remediation, and drawing up a recommendation for termination. The advice is summarized in three steps: (1) develop a sound evaluation system, (2) provide administrative inservice on evaluation policies and the proposed documentation system, and (3) apply in a way that preserves the personal dignity of the employees.

Frels, Kelly and Cooper, Timothy T. (1982). A documentation system for teacher improvement or termination. Practical concise guide for legal consideration in teacher evaluation. (Document No. EAQ15575). NOLPE Mini-Monograph. 24 p. (ERIC Document Reproduction Service No. ED228725).

Provided here is advice to principals concerning the documentation that should be used in the teacher evaluation process and presented at a hearing to support a recommendation for termination. The booklet presents a system of documentation intended to be simple but effective and which can be used in conjuction with any school district's evaluation system or contractual requirements. The documentation system is based on the value of communication, and its goal is to humanize the evaluation and documentation process. Explained in the booklet are several types of written memoranda and when they should be used: memoranda to the principal's file (for recording small infractions or deviations), specific incident memoranda (for recording conferences or significant events), summary memoranda (for recording conferences dealing with several incidents, problems, or deficiences), and visitation memoranda (for recording observations of teacher on-the-job

performance). The assessment or evaluation document is also discussed. Observations on documentation in general and a section on evaluating termination options conclude the booklet. Appendixes contain sample memoranda.

Mertens, Sally K. and Yarger, Sam J. (1979). Documenting success: a guidebook for teacher centers. (Document No. SP014736). Albany, New York: New York State Education Department, 37 p. (ERIC Document Reproduction Service No. ED175853).

The primary purpose of the documentation operations suggested by the National Teacher Centers Program is to aid individual projects in gaining a more profound and analytical understanding of what is happening within their own centers. The mechanism for this is the development of a massive information storage system, based on data collected from separate agencies. This guidebook outlines a strategy for the implementation of a recording process for programs without diverting great amounts of staff time and/or budget. The suggested documentation techniques can be easily incorporated into a program's prexisting activities. The importance of a favorable climate (in terms of staff attitude), pitfalls to documentation, and sample research instruments are discussed. Methods for analyzing data reaped from the informational system and the program documentation are included.

McLarty, Joyce and others (April, 1985). Using multiple data sources in teacher evaluation. (Document No. TM850556). 21 p. (ERIC Document Reproduction Service No. ED263122).

The Tennesse Career Ladder Evaluation system, which uses a multiple data source approach to assessment, was developed to identify excellent teachers. It is used to determine whether or not the teacher should receive increasing career benefits. Evaluation data are generated by three evaluators, the teacher candidate, the school principal, three peer teachers, and twenty or more students. Extensive review was used in generating six categories of teacher competency: planning; instructional strategies; evaluation, classroom management; leadership; and effective communication. Items were field tested in different types of data collection: classroom observation; interviews with teachers and principals; questionnaires for students and peer teachers; portfolio rating; a written test; and evaluator judgement. Results from each instrument were analyzed separately and then reanalyzed as part of the set. Intercorrelations were low, confirming that each source measured different viewpoints or aspects of performance. The resulting scores were reduced to a smaller number of scores, and then weighted. It was concluded that the multiple data source system was difficult and expensive to build, relatively inflexible, and complicated to explain. However, it provided a thorough and equitable evaluation, was relatively stable, and was logical.

Newcomb, B.J. and Murphy, P.A. (Sept., 1979). The curricular vitae -- what it is and what it is not. (Document NO. CE509047). Nursing Outlook. 27:9, pp. 580-583 (ERIC Document Reproduction Service No. EJ215543).

Presents guidelines to nurses for preparing curriculum vitae, narrower and more stylized than resumes, which outline qualifications for academic positions, committees, and honors. Lists categories and describes content of information to be included on a vita (a sample is shown) and notes what to avoid.

New Jersey Teachers' Salary Survey 1982-1983. (Document No. TM507751). New Jersey Education Association Review. 56:4, pp. 27-38 (ERIC Document Reproduction Service No. EJ275552).

Current data prepared by the New Jersey Education Association presents public schools' salary levels for each municipality in the state. Local degree requirements, minimum and maximum salaries for bachelors' and masters' degrees at four-, five-, and six-year training levels and increments are provided.

Perrone, Vito (April, 1977). Documentation: a process for classroom/program evaluation and personal/professional learning. Insights. 9:7, pp. 2-9.

Documentation of classroom processes and interaction plays two roles; it is helpful in classroom/program evaluation, and in the continuing personal and professional growth of the teacher. Documentation in the evaluation context is not primarily concerned with narrow objectives or with standardized tests, but rather seeks to determine meaning over time by recording the process of learning, the context of learning, and the content of learning. Documentation, in this respect, provides a comprehensive written record of the teaching/learning process which can be used in conjunction with other instruments for forming evaluative conclusions. As a process for personal/professional learning, documentation enables the teacher to analyze, reflect upon, and articulate basic assumptions about teaching, about teacher behavior in relation to students, parents, and colleagues, and about personal progress as a teacher. These two functions of documentation are not separate and distinct; the more collaborative they become, the more they become a base for interaction with others who are also thinking about their own teaching.

Scheyer, Patricia T. and Stake, Robert E. (1976). A Program's Self-Evaluation Portfolio. Urbana, Il.: University of Illinois.

Stedman, Carlton H. (1981). Credentialling: A Protean Model from other Professions. (Document No. SP023495). 13 p. (ERIC Document Reproduction Service No. ED238852).

Teacher education is faced with critical problems, including establishing its own credibility. Once assumption often made is that other professions do a more creditable job of monitoring those who enter their ranks and that teacher education might profit by emulating those groups who have been presumably more successful. The concept of relicensing deals with attempts to gain assurance that a practitioner has maintained a level of continuing competence. Of the procedures used in evaluating continuing competence in professions, two have received the bulk of attention: examinations, including clinical evaluations; and continuing education. Most professions have judged written examinations as inadequate for providing assurance to the public regarding continuing competence. Some form of on-the-job assessment has been recommended as a more logical, productive way of measuring competence. Reminders for teacher education include: (1) Licensing practices need constant reviewing and improvement; (2) No single test should be used to measure competence; (3) Validity studies are needed to establich competency measures' credibility; (4) Continuing education falls short of assuring practitioners' continuing competency; and (5) Viable models of cooperative efforts should be studied.

Terry, Gwenith L. and Eade, Gordon E. (1983). The Portfolio Process: New Roles for Meeting challenges in Professional Development. (Document No. SP022061). 12 p. (ERIC Document Reproduction Service No. ED229342).

The Portfolio Process, piloted by the University of West Florida and recently mandated by the Florida State Board of Education as the central element of the Florida Beginning Teacher Program, is a model for professional development of teaching competence. The portfolio is both a product and a process. As a product, the portfolio is a personalized compilation of information from multiple sources. Each piece of data included is placed in the notebook as a means of representing the individual's teaching competency and/or work toward increasing competency. The portfolio as a process includes three stages within a given cycle: (1) collection of data to support teaching competence; (2) review and analysis of data by a support team; and (3) commitment to a plan of action by the teacher. During the data collection stage of the process, each participant analyzes his own strengths and needs in each competency area, makes decisions concerning where and how the work begins, and collects data from various sources. During the second stage, the teacher presents evidence of competence to a support team who, in conferences with the teacher, reviews and analyzes the data. The final stage involves the teacher and support team agreeing on a plan of action for the teacher. This three-stage process is repeated until the desired level of competence is reached.

Terry, G.L. and Eade, G.E. (1979). A portfolio approach to professional development of the in-service and pre-service teacher. A paper presented at National Council of States on Inservice Education. Hollywood, Fl.: National Inservice Workshop, 11 p.

A description of the portfolio used by the Early Childhood faculty at the University of West Florida for both preservice and inservice teachers. The format of the portfolio specifies ten competencies and requires the individual teacher to assess and document strengths and needs in each competency.

Wilson, John D. (Dec., 1971). Data Processing at North Carolina Department of Public Instruction Division of Teacher Certification. (Document No. EA004386). 41 p. (ERIC Document Reproduction Service No. ED066791).

This document describes a data processing system now being implemented that is designed to provide rapid retrieval of information concerning a certificate holder's status. A brief history of past and present methods of teacher certification processing is included. These methods are then compared to the system currently being implemented. The new system provides the capability of collecting substantially more data and of storing and retrieving it in a meaningful report format. According to the author, the primary benefits of this sytem over the old are the ease with which inquiries can be answered, the more rigorous editing capabilities, and the faster computer service. Information on the costs, the procedures, and the problems of implementing the system and some indications about the directions of future changes in recordkeeping are included. Numerous charts and sample forms conclude the report.

Zenoff, Elyce H. and Barron, J.A. (Sept., 1983). So you want to hire a law professor? Journal of Legal Education. 33:3, pp. 492-510.

Problems in evaluating and predicting the competence of faculty applicants are discussed. Specific recommendations are made for improving the selection process, including creating an applicant "scorecard" of credentials, and improving professional association services and publications related to placement.

Beginning Teacher Portfolio Study

At this point, it should be noted that information is included in the appendices for variables represented in only 30 percent or more cases. The following variables were represented in fewer than 5 percent of the portfolios reviewed: type of certification (whether state-approved course,

alternative certification, etc.), the department or college degree, the college major, the date graduated from college, peer teacher's prior experience in that capacity and information regarding follow-up formative observations. In addition, two important categories: the teacher's prior experience and the subject area of certification for the peer teacher both were represented by information in fewer than 40 percent of the portfolios reviewed.

Table 23

Beginning Teacher Temporary Certification

Title	State ID	Cases N
Art Education (K-12)	114	2
Biology (7-12)	116	1
Business Education (7-12)	117	1
Chemistry (7-12)	118	1
Early Childhood (N-K)	122	1
Early Childhood/Elem. Ed.	123	6
Elementary Education (1-6)	131	2
Elem. Ed./ Mental Retrd.	134	1
Elem Ed./Mathematics	138	1
Elem. Ed./Early Child/ Mid. Sch. Soc. Std. (K-MS)	146	1
French (7-12)	161	1
Industrial Arts (7-12)	180	2
Mathematics (7-9)	190	1
Mathematics (7-12)	191	1
Math/Science (7-9)	194	1
Math/Science (7-12)	195	2
Math/Science (MS, 7-9)	196	1
Music Education (K-12)	202	1
Physical Education (K-12)	206	4
Spanish (7-12)	219	1
Social Studies (7-9)	225	2
Social Studies (7-12	226	1
Vocational Home Economics	236	2
Vocational Agriculture	238	1
	Total	40

valid cases ---------- 40 missing cases ---------- 93

Table 24

Peer Teacher Subject Area

Subject Area	Cases N	Pct.%
Language Arts	4	3.0
Foreign Language	2	1.5
Social Sciences/History	3	2.3
Mathematics	2	1.5
Science	3	2.3
PE or ROTC	1	0.8
Industrial Education	1	0.8
Home Economics	2	1.5
Music	1	0.8
Exceptional Child Educ.	3	2.3
Elementary Education	10	7.5
not given	101	75.9
Total	133	100.0

valid cases ---------- 32 missing cases ---------- 101

Table 25

Types of Documentation Found in Portfolios

Type of Documentation	Documented Competencies N	Pct.%
Cognitive Exam	116	3.2
Work Samples	738	20.5
Summative Observation	693	19.3
Formative Observation	1011	28.1
Interviews	394	10.9
Exhibits by Teacher	392	10.9
Combination Work & Intv.	245	6.8
Total Responses	3589	100.0

valid cases ---------- 118 missing cases ---------- 15

Relationship Between Screening Observation and PDPs

Due to the weak nature of the data involved in this appendix, the mapping of ineffective indicators to PDP assignments is shown in different fashions in Tables 26 and 27, and Figure 2. These tables represent a synthesis of 39 separate tables to increase comprehension. For this reason some of the numbers appear not to match properly. In cross-tabulation procedures, if a case is missing for one variable but not the other, the case becomes a missing case for both. Thus, raw frequency tables shown in other parts of this report will contain more cases than the following tables. In several instances, more than 100 percent of cases are represented in the PDP domain assignments. This occurs because some portfolios contain domain assignments relevant to a specific indicator for other reasons than that indicator's presence on the screening observation (for example, Domain 3 assignments may relate to several different indicators).

Table 26 contains a complete cross-tabulation of each ineffective indicator by each PDP assignment for all 46 domain documented cases. Table 27 contains the same basic information with percentages of the total sample (n = 46) instead of raw frequencies. Table 22 in Chapter 1 investigates the appropriateness of PDP remediation by including only percentages of relevant domains for a specific ineffective indicator.

Table 26 shows that for primary remediation domains, for all ineffective indicators except general praise (33%) and no homework-feedback (50%), two thirds or more of the cases having an ineffective indicator also show PDP remediation activities in the appropriate domain (Domain 3 for Domain 3 behaviors, etc.). Other domains, primarily Planning and Communication (1.0 and 5.0), are represented in from 38 to 100 percent of the PDPs.

The most commonly occurring ineffective indicators for this cross-tabulation closely match those found in the entire file (Table 22). Figure 2 clearly shows the continued emphasis on Domain 3 behaviors.

Table 26

Comparison of Ineffective Indicators to Recommended Domain

Domain	Ineffective Indicator	Total Documented Cases (n = 46) Including at Least One Ineffective Indicator on Screening Observation: Cases Having Indicator N	Assigned Domain in PDP: 1 N	2 N	3 N	4 N	5 N	6 N
3	Delays	11	10	*17	15	5	10	5
3	Disorganized	9	6	14	11	3	5	2
3	Unrelated talk/activity	9	3	6	6	1	3	-
3	Multiple questions	1	1	2	2	1	3	1
3	Non-academic questions	13	7	15	11	2	5	3
3	General praise	15	9	9	5	1	6	2
3	No homework/feedback	2	-	-	1	-	-	-
3	Remains at desk	4	5	7	4	2	5	1
4	Treats concepts poorly	5	7	8	7	4	4	4
4	Inadequate cause/effect	1	3	4	4	2	3	1

* This can be as many as 46, since PDPs can contain up to 6 domain assignments.

Table 27

Ineffective Indicators and Domain Assignment - Total Pcts

Ineffective Indicator	% Cases Having Indicator	Total Documented Cases Including at Least One Ineffective Indicator on Screening Observation: Domain Percent of Total (n = 46) 1 %	2 %	3 %	4 %	5 %	6 %
Delays	24	22	37	33	11	22	4
Disorganized	20	13	30	24	7	11	4
Unrelated talk-activity	20	7	13	13	2	7	-
Multiple questions	2	2	4	4	2	7	2
Non-academic questions	28	15	33	24	4	11	7
General praise	33	20	20	11	2	13	4
No homework-feedback	4	4	-	-	2	-	-
Remain at desk	9	11	15	9	4	11	2
Treats concept poorly	1	15	17	15	9	9	9
Poor cause-effect	2	7	9	9	4	7	2

Figure 2

Percentage of Ineffective Indicators

Indicator	Domain N	All Cases %	Cross-Map Cases %
Delays	3	31	24
Disorganized	3	23	20
Unrelated Talk/Activity	3	18	20
Non-Academic Questions	3	29	28
General Praise	3	26	33
Remains at Desk	3	13	9
Treats Concepts Poorly	4	16	11

Name of Reviewer _____

Number of Minutes for Review _____

Instruction for Reviewers of B.T.P. Portfolio Study

1. Introduction

This is an objective study of portfolios that represent a sampling of districts. The project report will not identify the teachers, the schools, or the data with specific districts. The data will be compiled and treated as group data.

2. Directions

a) On the cover of each instrument write your name and the amount of time you spent on the documentation of the portfolio.

b) Use marks wherever possible such as , or supply numbers such as in the case of dates.

c) There are 7 parts as follows:

1.0 and 2.0 are identifying information for B.T. and the team

3.0 = observations

4.0 = Professional Development Plan

5.0 = Competencies

6.0 = Conferences

7.0 = Comments by the Reviewer

d) If the B.T. has supplied exhibits, check them in the appropriate box if the documentation summary cited the exhibits as evidence. If the exhibits were not cited, ignore them. Do not analyze or evaluate the exhibits.

3. REMINDER: Collect the xeroxed copies of the diagnostic and final summative instruments from the district coordinator.

What's In a Portfolio?

Chapter 6A-5.75 B.T.P.

The portfolio shall include, but is not limited to,

1) The Professional Development Plan, including 2 summative evaluations, at least 3 formative evaluation conferences and criteria for assessment.

2) The formative information

3) Exhibits of successful performance, included by the B.T. if desired.

BEGINNING TEACHER PROGRAM

Data Collection Instrument or Portfolio Evaluation

1.0 Identification of Beginning Teacher

1.1 Teacher's name _____

1.2 Social security number _____

1.3 District or non public association name _____

1.4 School _____

1.5 Temporary certification

1.5.1 Area _____

1.5.2 Number _____

1.6 Type of certification program (check one)

1.6.1 State-approved program _____

1.6.2 Course-by-course _____

1.6.3 Non degree vocational _____

1.6.4 Alternative certification _____

1.7 Institution from which graduated _____

1.7.1 Department or college _____

1.7.2 Program or major _____

1.7.3 Date of graduation _____

1.8 Teaching assignment

1.8.1 Subject (s) _____

1.8.2 Grade level (s) _____

1.9 Employment date _____

1.10 Number of years teaching experience prior to current year _____

2.0 Identification of Support Staff

2.1 Building administrator name (BLA) _____

2.1.1 Position _____

2.2 Peer teacher name (PT) _____

2.2.1 Teaching assignment/position _____

2.2.2 Check one: Experienced in this role: yes _____ no _____

2.3 Other professional educator name (OE) _____

2.3.1 Teaching assignment/position _____

2.3.2 Work location _____

3.0 Observations of B. T.

Write in the boxes the dates the instruments were used.

Type of Instrument	Diagnostic Observation	Intermediate or Follow-up Observation	Final Observation	Give the Total Number Of Completed Instruments In Portfolio
FPMS Summative Instrument				
Domain 1.0 Formative Instrument				
Domain 2.0 Formative Instrument				
Domain 3.0 Formative Instrument				
Domain 4.0 Formative Instrument				
Domain 5.0 Formative Instrument				
Domain 6.0 Formative Instrument				
Others, Specify				

3.1

Number of Ineffective Behaviors Identified on the First Diagnostic Summative Instrument

Date	Check Observer BLA/PT/OE	List Indicator Number(s)	Give Number of Tallies

If another diagnostic summative observation was made during the first 45 days list the ineffective behavior identified.

Date	Check Observer BLA/PT/OE	List Indicator Number(s)	Give Number of Tallies

4.0 Professional Development Plan (PDP)

Check whether documentation is recorded by competencies or domains:

☐ Competencies ☐ Domains

What type of inservice activities are prescribed?

COMPETENCY	Readings	Observations of Other Teachers	View Videotapes	Self Instructional Materials	Coaching	Workshops	Other	Is Follow-up Time Clearly Stated (Name If Yes)	How is Assessment Activity to be Determined? (Name all that Appear)				When is the Activity		
									Interview	Observation	Videotape	Work Sample	BLA	PY	
1. Oral Communication															
2. Written Communication															
3. Listening Comprehension															
4. Comprehension of Professional Reading															
5. Fundamental Mathematical Concepts															
6. Student Growth and Development															
7. Diagnosis of Entry Level Knowledge/Skills															
8. Identification of Long Range Goals															
9. Construct/Sequence Short Range Objectives															
10. Select/Adapt/Develop Instructional Materials															
11. Select/Develop Sequence Activities															
12. Establish Rapport with Students															
13. Presents Directions for an Instructional Activity															
14. Constructs/Assembles Classroom Test															
15. Establishes Routines for Care and Use of Materials															
16. Formulates a Standard for Student Behavior															
17. Uses Behavior Management Techniques															
18. Establishes a Recordkeeping System															

167

BEST COPY AVAILABLE

4.0 Continued

COMPETENCY	Readings	Observations of Other Teachers	View Videotapes	Self Instructional Materials	Coaching	Workshops	Other	If Follow-up Time Clearly Stated (Name if Yes)	How is Assessment Activity to be Determined? (Name all that Apply)				Who is to Assess the Activity?		
									Interview	Observation	Videotape	Self Report	DLA	PT	CE
30. Demonstrates Feeling for Dignity and Worth of Others															
31. Assist Students with Positive Self Concept Development															
32. Facilitates Positive Interaction Among Peers															
33. Facilitates Values Clarification															
34. Recognizes/Meets Needs of Exceptional Students															
Domain 1															
Domain 2															
Domain 3															
Domain 4															
Domain 5															
Domain 6															

BEST COPY AVAILABLE

5.0 Documentation of Competence

Check the mean(s) by which each competency was certified by the support staff.

COMPETENCY	Cognitive Exam	Analysis of Work Samples	Summative Classroom Observation	Formative Classroom Observation	Interview	Exhibits Added by B.T.: Specify	Is Documentary Evidence Present In Portfolio
1. Oral Communication							
2. Written Communication							
3. Listening Comprehension							
4. Comprehension of Professional Reading							
5. Fundamental Mathematical Concepts							
6/19. Student Growth and Development							
7. Diagnosis of Entry Level Knowledge/Skills							
8. Identification of Long-range Goals							
9. Construct/Sequence Short Range Objectives							
10. Select/Adapt/Develop Instructional Materials							
11. Select/Develop Sequence Activities							

5.0 Continued

COMPETENCY	Cognitive Exam	Analysis of Work Samples	Summative Classroom Observation	Formative Classroom Observation	Interview	Exhibits Added by B.T.: Specify	Is Documentary Evidence Present In Portfolio
12. Establish Rapport with Students							
13. Presents Directions for an Instructional Activity							
14. Construct/Assembles Classroom Test							
15. Establishes Routines for Care and Use of Materials							
16. Formulates a Standard for Student Behavior							
17. Uses Behavior Management Techniques							
18. Establishes a Recordkeeping System							
20. Demonstrates Feeling for Dignity and Worth of Others							
21. Assist Students with Positive Self Concept Development							
22. Facilitates Positive Interaction Among Peers							
23. Facilitates Values Clarification							
24. Recognizes/Meets Needs of Exceptional Students							

6.0 Record of Conferences

TYPE OF CONFERENCE				WHO ATTENDED			TOTAL NUMBER OF CONFERENCES		
Pre Observation Date	Post Observation Date	Follow Up Date	Final Assessment Date	BLA	PT	OE	Formative	Summative	Other

KEY: BLA = Building Level Administrator
PT = Peer Teacher
OE = Other Educator
BT = Beginning Teacher

7.0 Comments:

FLORIDA ESSENTIAL COMPETENCIES
SAMPLE METHODS OF DEMONSTRATION

COMPETENCY		CLASSROOM DEMONSTRATION			COGNITIVE EXAMINATION	OTHER DATA SOURCES
No.	Description	Summative Item(s)	Formative Domain	Formative Concepts		
1.	Oral Communication	17	5	5.1		
2.	Written Communication				Florida Teacher Certification Examination (FTCE)	Domain 1.0 — Concepts 1.1 1.2 1.3 1.4 1.5
3.	Listening Comprehension		3 6	3.4 3.5 6.2 6.3		
4.	Comprehension of Professional Reading				FTCE	Successful Completion of any FPMS Learning Package.
5.	Comprehend/work with fundamental Mathematical Concepts				FTCE CLAST	

COMPETENCY		CLASSROOM DESCRIPTION			COGNITIVE EXAMINATION	OTHER DATA SOURCES
No.	Description	Summative Item(s)	Formative Domain	Formative Concepts		
6/19	Student Growth and Development				FICE	
7.	Diagnosis of Entry Level Knowledge/ Skills					Domain 1.0 Concept 1.5
8.	Identification of Long Range Subject Area Goals					1. A teacher prepared list of unit, semester or year-long goals.
9.	Construct and Sequence Short Range Objectives		3.0	3.3 3.6 6.1		Domain 1.0 Concept 1.1 2. A teacher prepared learning packet for students.
10.	Select/Adapt/Develop Instructional Materials for Objectives					Domain 1.0 Concept 1.2 2. A teacher prepared learning center.
11.	Select/Develop/ Sequence Activities for Objectives					Domain 1.0 Concept 1.3 2. A teacher prepared learning center.

COMPETENCY		CLASSROOM DESCRIPTION			COGNITIVE EXAMINATION	OTHER DATA SOURCES
No.	DESCRIPTION	SUMMATIVE ITEM(S)	FORMATIVE DOMAIN	FORMATIVE CONCEPTS		
12.	Establish Rapport with Students	15 16 19	2.0 5.0 6.0	2.8 6.1 5.2 6.2 5.3 6.3 5.4 5.5		
13.	Presents Directions for Instructional Activity	3 9	3.0	3.3 3.6		
14.	Consturcts/Assembles Classroom Test					1. Sample teacher made tests. 2. Satisfactory completion of the learning peackage Preparing Tests.
15.	Established Routines for care and use of Materials	7	1.0	1.1		
16.	Formulates a Standard for Student Behavior		2.0	2.1		
17.	Uses Behavior Management Techniques	20 21	2.0	2.2 2.6 2.1 2.8 2.4		

BEST COPY AVAILABLE

COMPETENCY		CLASSROOM DESCRIPTION			COGNITIVE EXAMINATION	OTHER DATA SOURCES
No.	DESCRIPTION	SUMMATIVE ITEM(S)	FORMATIVE DOMAIN	FORMATIVE CONCEPTS		
18.	Establishes a Recordkeeping System					1. Grade Book Records. 2. Basic Skills Checklist Records. 3. Pupils Progression Plan Records of Student Progress. 4. Individual Education Plan. 5. Prep Strategies.
20.	Demonstrates Feeling for Dignity and Worth of Others	6 7	3.0 5.0 6.0	3.4 5.5 6.1 6.3		
21.	Assists Students with Positive Self Concept Development	6 7 16 20	2.0 3.0 5.0 6.0	2.2 6.3 2.8 3.4 5.5 6.1		
22.	Facilitates Positive Interaction Among Peers	6	3.0	3.3 3.4 3.5		
23.	Assists Students in Acquisition of Decision-making Skills	14	2.0 4.0	2.1 4.4		
24.	Recognizes Needs of Exceptional Students				FICE	Domain: 1.0 Concepts: 1.2, 1.3, 1.5

175

COMPETENCY		CLASSROOM DESCRIPTION			COGNITIVE EXAMINATION	OTHER SOURCES
No.	DESCRIPTION	SUMMAVTIVE ITEM(S)	FORMATIVE DOMAIN	CONCEPTS		
25.	Demonstrates Questioning Techniques	5	3.0 5.0	3.3 5.1		
26.	Promotes Learning/ Retention through Practice	8 9	3.0	3.3 3.6		
27.	Encourages Student Participation that Maintains Academic Focus	3 5 6 7	3.0	3.3 3.4 3.5		
28.	Uses Feedback Procedures	6 7 9	3.0 6.0	3.4 3.5 3.6 6.3		
29.	Conducts Subject Matter Review	4	3.0 6.0	3.2 6.1		
30.	Uses Class Time Efficiently	1 20 2 21 1	2.0 3.0	2.2 2.7 2.3 3.1 2.4 2.6		

COMPETENCY		CLASSROOM DESCRIPTION			COGNITIVE EXAMINATION	OTHER DATA SOURCES
No.	DESCRIPTION	SUMMATIVE ITEM(S)	FORMATIVE DOMAIN	FORMATIVE CONCEPTS		
31.	Presents Knowledge: Concepts, Laws, and Rules	11 12 13	4.0	4.1 4.2 4.3		
32.	Demonstrates Control of Vocal Expression	16 17 18	5.0	5.3 5.4		
33.	Uses Body Behavior that Expresses Interest	19	5.0	5.5		
34.	Utilizes Effective Examination Procedures		6.0	6.1 6.2 6.3		
35.	Recognizes Severe Emotional Distress and Utilizes Crisis Intervention Techniques					1. Record of Appropriate Referral. 2. Successful Completion of Approved MIP Workshop. 3. Written response to a standard case Study.

State of Florida
Department of Education
Tallahassee, Florida
Ralph D. Turlington, Commissioner
Affirmative action/equal opportunity employer

FLORIDA: A STATE OF EDUCATIONAL DISTINCTION. "On a statewide average, educational achievement in the State of Florida will equal that of the upper quartile of states within five years, as indicated by commonly accepted criteria of attainment."

APPENDIX B

APPENDIX B

Predictive Validity Study Designs and Findings: Simple Correlations Across Studies

Tables 11 and 12 showed the simple Pearson product moment correlations between variables based upon teacher/class means for Studies A and B, and teacher means for Study C (there were seven more classes than teachers in Study C). Dependent variables included residualized gain scores (RGS), raw posttest scores for the average student stratum (ASP), and student task engagement scores. Independent variables included FPMS total scores and student task engagement scores.

Tables 13 and 14 showed approximate correlations that could associate with the combined Z scores reported in Tables 2 and 4 in Chapter 4 of the report. The approximate r reported in these tables is derived from the probabilities associated with an average combined Z of residualized gain scores and average stratum posttest scores. The average r reported in these tables is a simple average of the correlation coefficients reported in Tables 11 and 12. The simple average, does not take into account differences in the sample sizes for various studies.

TABLE 11
Correlations Between FPMS Scores and Dependent Variables

Study			Residualized Gain Scores	Average Stratum	Task Engagement
Study A -	Math	r=	.57	.31	.42
			n=20	n=17	n=21
			p<.005*	p<.111	p<.029
			z=2.58**	z=1.22	z=1.90
	Soc Std.	r=	.22	.28	.44
			n=29	n=30	n=31
			p<.131	p<.064	p<.006
			z=1.13	z=1.52	z=2.51
	Both	r=	.33	.28	.43
			n=49	n=47	n=52
			p<.011	p<.029	p<.001
			z=2.32	z=1.90	z=3.10
Study B	Am. Hist.	r=	.13	.095	.65
			n=38	n=38	n=38
			p<.220	p<.285	p<.0001
			z=0.77	z=0.56	z=3.70

TABLE 11 (continued)
Correlations Between FPMS Scores and Dependent Variables

Study		Residualized Gain Scores	Average Stratum	Task Engagement
Study C Math	r=	.109	.346	
		n=25	n=25	
		p<.302	p<.045	
		z=0.52	z=1.69	

* one-tailed P value
** one-tailed Z value

TABLE 12
Correlations Between Task Engagement and Achievement

Study		Residualized Gain Scores	Average Stratum
Study A - Math	r=	.54	.16
		n=19	n=16
		p<.008	p<.39
		z=2.41	z=0.27
Soc Std.	r=	.43	.39
		n=29	n=30
		p<.011	p<.017
		z=2.29	z=2.12
Both	r=	.47	.29
		n=48	n=46
		p<.0001	p<.027
		z=3.70	z=1.92
Study B Am. Hist.	r=	.28	.27
		n=38	n=38
		p<.043	p<.052
		z=1.71	z=1.62

* one-tailed P value
** one-tailed Z value

TABLE 13

Approximate Correlations Associating with Combined Z Scores Between FPMS Scores and Achievement

Grouping	N of Studies	N of Cases	Avg Z	Avg P	Aprox+ R	Avg+ R
All Studies	4	111	2.50*	.006**	.25	.26
Mathematics	2	43	2.10	.018	.33	.33
Social Std.	2	66	1.40	.081	.18	.18
Elementary	3	73	2.50	.006	.31	.31
Post-Elem.	1	38	.66	.255	---	.11

* one-tailed Z
** one-tailed P

\+ obtained by using Z probabilities
++ obtained by averaging correlation coefficients

TABLE 14

Approximate Correlations Associating with Combined Z Scores Between Task Engagement and Achievement

Grouping	N of studies	N of cases	Avg Z	Avg P	Aprox r	Avg r
All Studies	3	86	3.00*	.001**	.35	.35
Mathematics	1	19	1.34	.091	.38	.35
Social Std.	2	66	2.74	.031	.35	.34
Elementary	2	48	2.50	.006	.34	.38
Post-Elem.	1	38	1.66	.049	.26	.27

* one-tailed Z
** one-tailed P

Specification of Outliers

The following discussion details the rationale behind the elimination of two subjects from Study B data and one from Study C. The first individual was determined to be an outlier in several senses of the term.

1. The individual's FPMS score was 2.5 standard deviations below the sample mean.
2. The individual's task engagement score was over 3 standard deviations below the sample mean.
3. The relationship between the individual's ineffective scores and class posttest scores exhibited extremely unusual characteristics (Table 15).
4. The relationship between the individual's class pretest and posttest scores can only be described as anomalous since almost all low scoring pretest students ceilinged out, while three high scoring students scored in the 10th percentile on the posttest.
5. The relationship between the individual's engagement scores and class posttest scores was extremely unusual. With this individual included in the data, the simple correlation between task engagement and achievement was negative. With this individual eliminated from the data, the relationship between task engagement and achievement became positive and bordered closely on significance ($r=.263$, $P <= .056$).

For the above noted reasons, this person's scores were deleted from the data for these analyses. Table 15 plots the anomalous relationship between students' pre and posttest scores for this teacher.

A second outlier having a residualized gain score 3.608 standard deviations above the mean appeared in Study B data. Table 17 depicts the leverage this individual would apply to any correlations if his/her score were retained during analysis. This case was deleted from further analysis.

A single outlier having an extremely low average FPMS score was deleted from study B data. This person's score was 3.15 standard deviations below the mean and 2.09 standard deviations below the next lowest scoring teacher.

Due to the nature of both parametric and non-parametric statistical techniques, outliers such as the ones noted above can substantially alter the results of analysis. The question arises whether

the individuals exhibiting such results are truly part of the same population as other individuals, or are rare and unique cases whose influence merely confuses researchers attempting to reach "generalizeable" conclusions. Such conclusions can never apply in all cases, but hopefully they can be accurate for the majority of individuals. The following quote from Andrews et. al (1972) deals with generally accepted techniques applied by practical statisticians to robustify their results and conclusions.

> "A good statistician (or research worker) never uses the mean, even though he may honestly claim to do so. In reality, he will look at the data and set aside outliers and suspected outliers in order to treat them separately; then he will take the mean of the rest. ...And since such a rejection procedure will get rid of distant contamination, any such method will work in that it prevents the worst" (damage to results).

Obviously, in the case of correlation coefficients, where one deals with squared deviations, this process becomes even more important, as Wainer (1982) suggests.

TABLE 15

Plot of Pre and Posttests for Outlier - Study B
(8th grade American History – FPMS Score = 53.5)

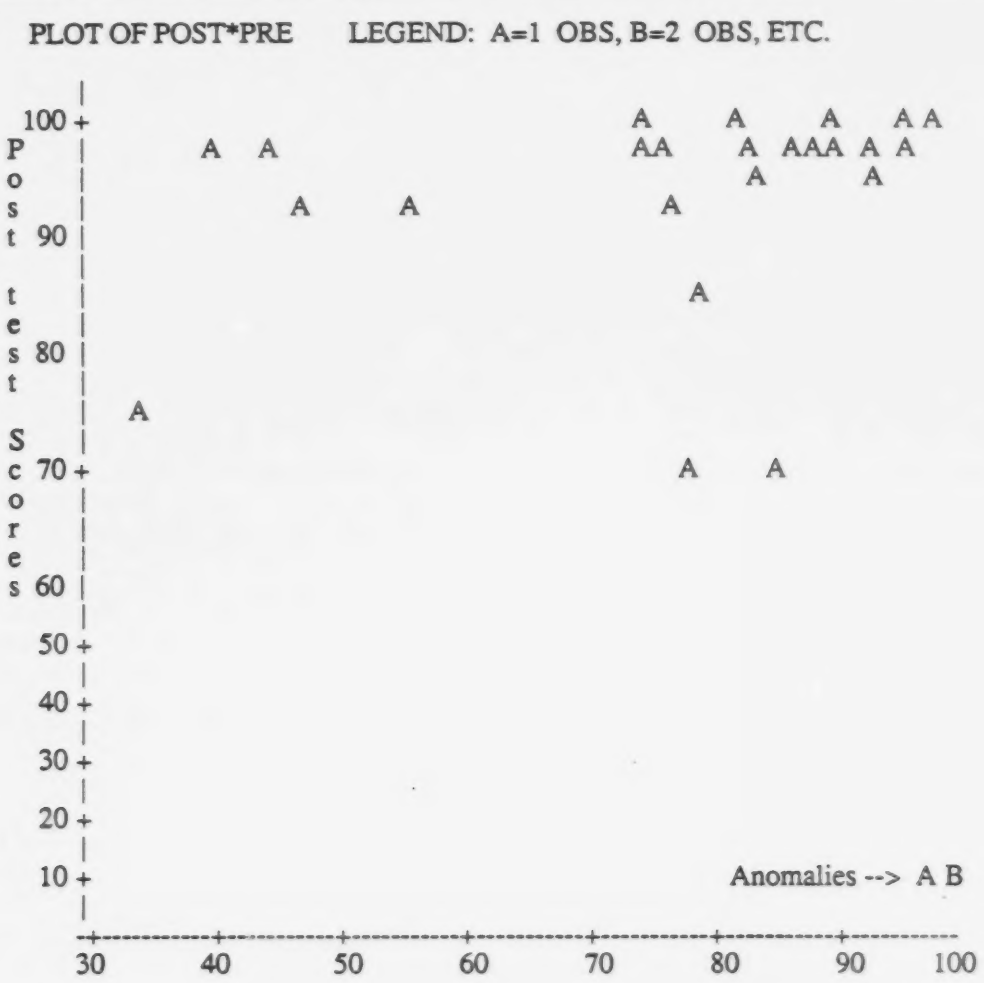


TABLE 16

Plot of FPMS Ineffective Scores and Class Posttest Means

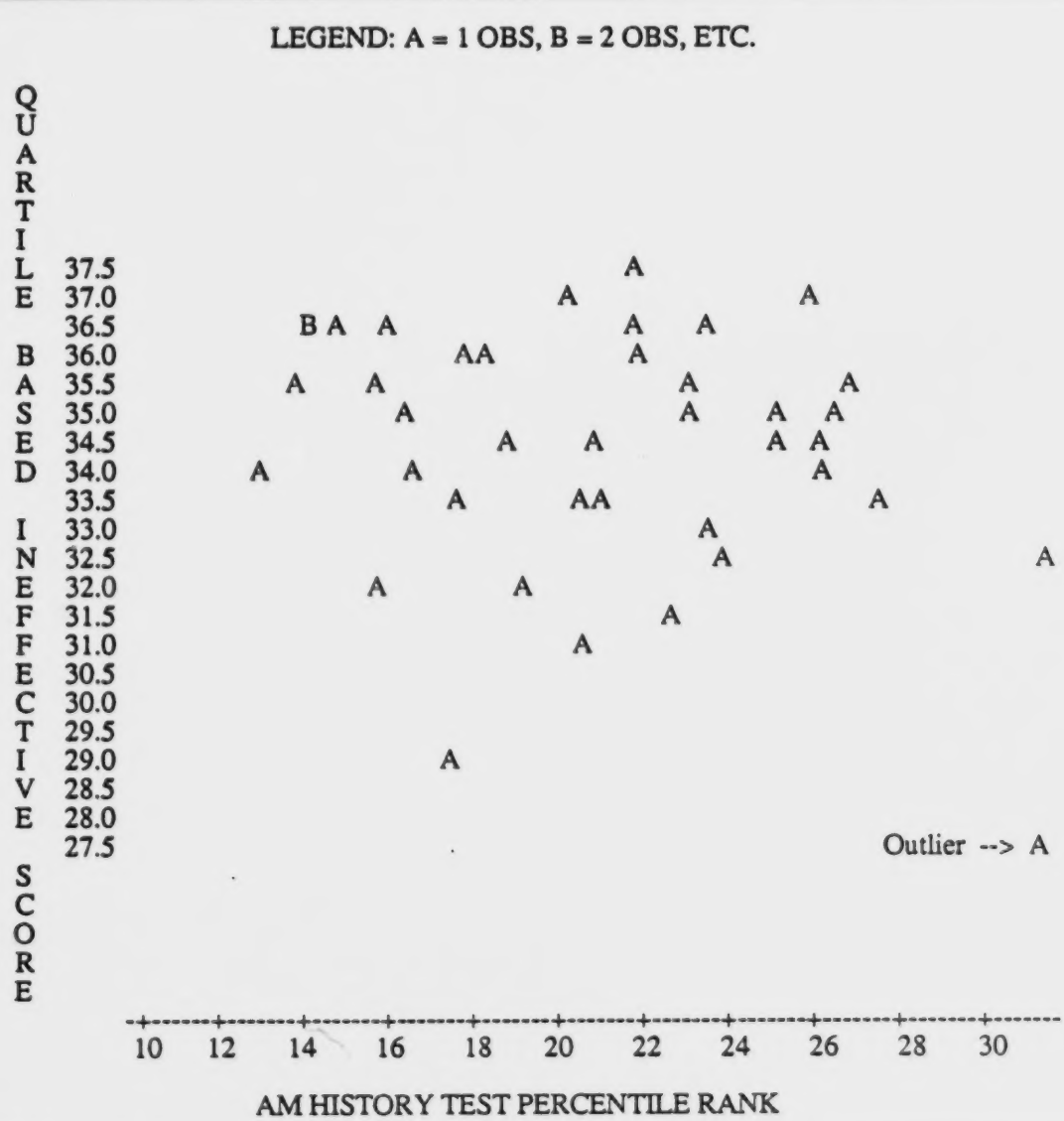


TABLE 17

Plot Of Residualized Scores And FPMS Scores Showing Outlier

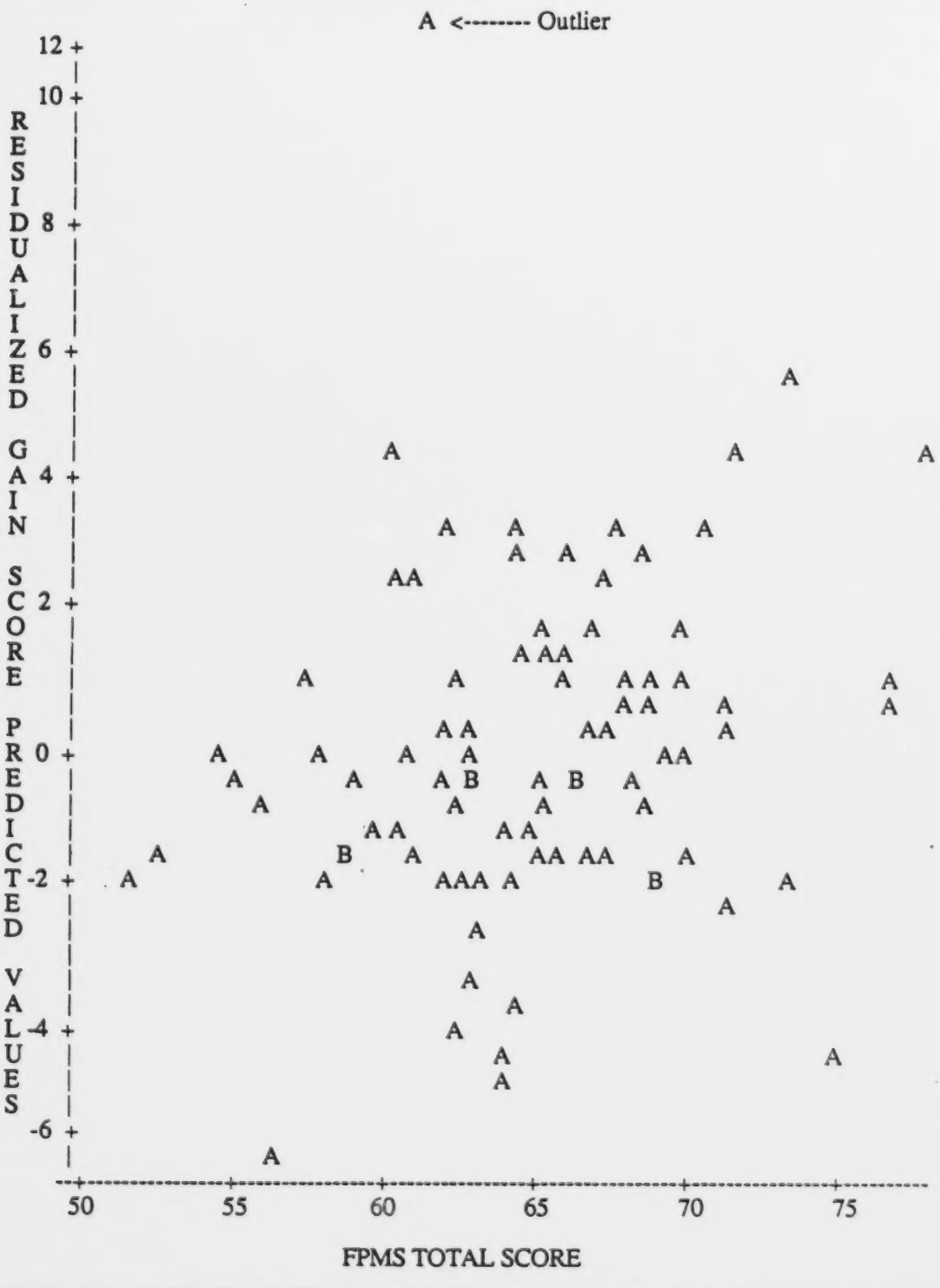


TABLE 18

Study A - FPMS Total Score Frequencies by Teacher

Score Code	Absolute Freq	Relative Freq (Pct)	Adjusted Freq (Pct)	Cum Freq (Pct)
56.50	1	1.7	1.9	1.9
61.00	2	3.4	3.8	5.7
63.50	2	3.4	3.8	9.4
64.00	2	3.4	3.8	13.2
64.50	1	1.7	1.9	15.1
65.00	1	1.7	1.9	17.0
66.00	4	6.9	7.5	24.5
67.00	1	1.7	1.9	26.4
67.50	2	3.4	3.8	30.2
68.00	3	5.2	5.7	35.8
68.50	1	1.7	1.9	37.7
69.00	2	3.4	3.8	41.5
69.50	2	3.4	3.8	45.3
70.50	2	3.4	3.8	49.1
71.00	2	3.4	3.8	52.8
71.50	1	1.7	1.9	54.7
72.00	1	1.7	1.9	56.6
72.50	3	5.2	5.7	62.3
73.00	3	5.2	5.7	67.9
74.00	1	1.7	1.9	69.8
74.50	4	6.9	7.5	77.4
75.00	1	1.7	1.9	79.2
75.50	5	8.6	9.4	88.7
77.00	2	3.4	3.8	92.5
79.00	2	3.4	3.8	96.2
81.00	2	3.4	3.8	100.0
-9.00	5	8.6	missing	100.0
Total	58	100.0	100.0	

mean	70.566	std err	0.731	median	70.875
mode	75.500	std dev	5.319	variance	28.289
kurtosis	-0.187	skewness	-0.227	range	24.500
minimum	56.500	maximum	81.000		

Valid Cases 53 Missing Cases 5

TABLE 19

Study A - Residualized Gain Score Frequencies

Score Code	Freq	Adj Pct	Cum Pct	Score Code	Freq	Adj Pct	Cum Pct	Score Code	Freq	Adj Pct	Cum Pct
-4.29	1	2	2	-1.21	1	2	37	1.26	1	2	71
-3.21	1	2	4	-0.90	1	2	39	1.32	1	2	73
-2.41	1	2	6	-0.83	1	2	41	1.37	1	2	76
-2.15	1	2	8	-0.74	1	2	43	1.38	1	2	78
-2.12	1	2	10	-0.36	1	2	45	1.59	1	2	80
-2.06	1	2	12	-0.28	1	2	47	1.84	1	2	82
-1.97	1	2	14	-0.23	1	2	49	1.88	1	2	84
-1.94	1	2	16	-0.15	1	2	51	1.94	1	2	86
-1.91	1	2	18	-0.08	1	2	53	2.50	1	2	88
-1.88	1	2	20	0.03	1	2	55	2.70	1	2	90
-1.70	1	2	22	0.36	1	2	57	2.81	1	2	92
-1.69	1	2	24	0.53	1	2	59	3.10	1	2	94
-1.62	1	2	27	0.59	1	2	61	3.19	1	2	96
-1.62	1	2	29	0.74	1	2	63	4.50	1	2	98
-1.56	1	2	31	0.96	1	2	65	5.71	1	2	100
-1.41	1	2	33	0.96	1	2	67				
-1.23	1	2	35	1.06	1	2	69				

MISSING DATA

CODE	FREQ	CODE	FREQ	CODE	FREQ
-9.00	9				

mean	0.056	std err	0.293	median	-0.151
mode	-4.292	std dev	2.051	variance	4.207
kurtosis	0.107	skewness	0.471	range	10.001
minimum	-4.292	maximum	5.710		

Valid Cases 49 Missing Cases 9

TABLE 20

STUDY A - Average Stratum Posttest Scores - ASP

Score Code	Freq	Adj Pct	Cum Pct	Score Code	Freq	Adj Pct	Cum Pct	Score Code	Freq	Adj Pct	Cum Pct
7.50	1	2	2	15.67	1	2	38	17.13	1	2	72
11.67	1	2	4	15.75	1	2	40	17.35	1	2	74
13.00	1	2	6	15.80	1	2	43	17.71	1	2	77
13.25	1	2	9	15.88	1	2	45	17.85	1	2	79
13.89	1	2	11	16.00	1	2	47	18.00	1	2	81
14.33	1	2	13	16.11	1	2	49	18.45	1	2	83
14.50	1	2	15	16.14	1	2	51	19.00	1	2	85
14.73	1	2	17	16.20	2	4	55	19.50	1	2	87
14.75	1	2	19	16.40	1	2	57	19.57	1	2	89
14.83	1	2	21	16.50	1	2	60	20.14	2	4	94
15.00	2	4	26	16.60	1	2	62	20.20	1	2	96
15.25	1	2	28	16.75	1	2	64	20.50	1	2	98
15.27	1	2	30	16.86	1	2	66	24.60	1	2	100
15.33	1	2	32	16.89	1	2	68				
15.50	2	4	36	17.13	1	2	70				

MISSING DATA

CODE	FREQ	CODE	FREQ	CODE	FREQ
-9.00	11				

mean	16.390	std err	0.386	median	1.143
mode	15.000	std dev	2.648	variance	7.014
kurtosis	3.234	skewness	-0.060	range	17.100
minimum	7.500	maximum	24.600		

Valid Cases 47 Missing Cases 11

TABLE 21

Study B - FPMS Total Score Frequencies By Teacher

Score Code	Freq	Adj Pct	Cum Pct	Score Code	Freq	Adj Pct	Cum Pct	Score Code	Freq	Adj Pct	Cum Pct
0.65	1	2	2	0.87	1	2	37	0.91	1	2	71
0.70	1	2	4	0.87	1	2	38	0.92	1	2	73
0.76	1	2	6	0.87	1	2	40	0.92	1	2	75
0.78	1	2	8	0.87	1	2	42	0.92	1	2	77
0.78	1	2	10	0.87	1	2	44	0.93	1	2	79
0.78	1	2	12	0.87	1	2	46	0.93	1	2	81
0.78	1	2	13	0.88	1	2	48	0.94	1	2	83
0.80	1	2	15	0.88	1	2	50	0.94	1	2	85
0.80	1	2	17	0.88	1	2	52	0.94	1	2	87
0.80	1	2	19	0.88	1	2	54	0.94	1	2	88
0.81	1	2	21	0.89	1	2	56	0.94	1	2	90
0.85	1	2	23	0.90	1	2	58	0.95	1	2	92
0.85	1	2	25	0.90	1	2	60	0.95	1	2	94
0.85	1	2	27	0.90	1	2	62	0.96	1	2	96
0.85	1	2	29	0.90	1	2	63	0.98	1	2	98
0.85	1	2	31	0.91	1	2	65	0.98	1	2	100
0.85	1	2	33	0.91	1	2	67				
0.86	1	2	35	0.91	1	2	69				

MISSING DATA

CODE	FREQ	CODE	FREQ	CODE	FREQ
-9.00	6				

mean	0.874	std err	0.009	median	0.880
mode	0.649	std dev	0.068	variance	0.005
kurtosis	1.526	skewness	-1.068	range	0.333
minimum	0.649	maximum	0.981		

valid cases 52 missing cases 6

TABLE 22

Study B - FPMS Total Score Frequencies By Teacher

Score Code	Absolute Freq	Relative Freq (Pct)	Adjusted Freq (Pct)	Cum Freq (Pct)
54.50	1	2.6	2.6	2.6
57.50	1	2.6	2.6	5.3
59.50	1	2.6	2.6	7.9
60.00	1	2.6	2.6	10.5
61.50	3	7.9	7.9	18.4
62.00	1	2.6	2.6	21.1
62.50	1	2.6	2.6	23.7
63.00	2	5.3	5.3	28.9
64.50	1	2.6	2.6	31.6
65.00	3	7.9	7.9	39.5
65.50	1	2.6	2.6	42.1
66.00	2	5.3	5.3	47.4
66.50	1	2.6	2.6	50.0
67.00	2	5.3	5.3	55.3
67.50	2	5.3	5.3	60.5
68.00	3	7.9	7.9	68.4
69.50	1	2.6	2.6	71.1
70.00	1	2.6	2.6	73.7
70.50	2	5.3	5.3	78.9
71.50	2	5.3	5.3	84.2
72.50	1	2.6	2.6	86.8
73.50	2	5.3	5.3	92.1
74.00	1	2.6	2.6	94.7
74.50	1	2.6	2.6	97.4
76.00	1	2.6	2.6	100.0
Total	38	100.0	100.0	

mean	66.592	std err	0.809	median	66.750
mode	61.500	std dev	4.989	variance	24.890
kurtosis	-0.277	skewness	-0.186	range	21.500
minimum	54.500	maximum	76.000		

valid cases 38 missing cases 0

TABLE 23

STUDY B - Residualized Gain Score (RGS) Frequencies

Score Code	Freq	Adj Pct	Cum Pct	Score Code	Freq	Adj Pct	Cum Pct	Score Code	Freq	Adj Pct	Cum Pct
-6.83	1	3	3	-1.05	1	3	37	0.81	1	3	71
-5.26	1	3	5	-0.76	1	3	39	1.00	1	3	74
-4.95	1	3	8	-0.57	1	3	42	1.18	1	3	76
-4.92	1	3	11	-0.56	1	3	45	1.49	1	3	79
-3.79	1	3	13	-0.54	1	3	47	1.75	1	3	82
-3.51	1	3	16	-0.53	1	3	50	2.48	1	3	84
-2.84	1	3	18	-0.49	1	3	53	2.56	1	3	87
-2.01	1	3	21	-0.40	1	3	55	2.68	1	3	89
-1.96	1	3	24	-0.10	1	3	58	3.13	1	3	92
-1.60	1	3	26	-0.08	1	3	61	3.20	1	3	95
-1.60	1	3	29	-0.07	1	3	63	4.45	1	3	97
-1.44	1	3	32	0.28	1	3	66	4.60	1	3	100
-1.27	1	3	34	0.44	1	3	68				

mean	-0.449	std err	0.430	median	-0.531
mode	-6.833	std dev	2.649	variance	7.016
kurtosis	0.059	skewness	-0.294	range	11.432
minimum	-6.833	maximum	4.599		

valid cases 38 missing cases 0

TABLE 24

STUDY B - Average Stratum Posttest (ASP) Frequencies

Score Code	Freq	Adj Pct	Cum Pct	Score Code	Freq	Adj Pct	Cum Pct	Score Code	Freq	Adj Pct	Cum Pct
13.40	1	3	3	18.79	1	3	37	22.94	2	5	74
14.10	1	3	5	18.90	1	3	39	23.27	1	3	76
14.78	1	3	8	19.07	1	3	42	23.31	1	3	79
15.04	1	3	11	20.20	1	3	45	23.33	1	3	82
15.35	1	3	13	20.33	2	5	50	23.38	1	3	84
16.00	1	3	16	20.88	1	3	53	24.17	1	3	87
16.47	1	3	18	21.10	1	3	55	24.83	1	3	89
16.89	1	3	21	21.12	1	3	58	25.00	1	3	92
17.43	1	3	24	21.79	1	3	61	25.13	1	3	95
18.33	1	3	26	21.85	1	3	63	25.67	1	3	97
18.63	1	3	29	22.00	1	3	66	26.67	1	3	100
18.67	2	5	34	22.56	1	3	68				

mean	20.350	atd err	0.568	median	20.344
mode	18.667	std dev	3.504	variance	12.280
kurtosis	-0.828	skewness	-0.229	range	13.267
minimum	13.400	maximum	26.667		

valid cases 38 missing cases 0

TABLE 25

Study C - FPMS Total Score Frequencies

Score Code	Absolute Freq	Relative Freq (Pct)	Adjusted Freq (Pct)	Cum Freq (Pct)
58.75	1	3.7	3.8	3.8
66.50	2	7.4	7.7	11.5
68.13	1	3.7	3.8	15.4
68.50	1	3.7	3.8	19.2
68.75	1	3.7	3.8	23.1
69.00	2	7.4	7.7	30.8
69.25	1	3.7	3.8	34.6
69.50	2	7.4	7.7	42.3
69.75	1	3.7	3.8	46.2
69.79	1	3.7	3.8	50.0
70.50	3	11.1	11.5	61.5
70.75	1	3.7	3.8	65.4
71.00	1	3.7	3.8	69.2
71.25	1	3.7	3.8	73.1
72.50	2	7.4	7.7	80.8
74.50	1	3.7	3.8	84.6
74.75	1	3.7	3.8	88.5
76.00	1	3.7	3.8	92.3
76.50	1	3.7	3.8	96.2
77.00	1	3.7	3.8	100.0
-9.00	1	3.7	missing	100.0
Total	27	100.0	100.0	

mean	70.429	std err	0.727	median	69.812
mode	70.500	std err	3.705	variance	13.728
kurtosis	3.079	skewness	-0.768	range	18.250
minimum	58.750	maximum	77.000		

valid cases 26 missing cases 1

TABLE 26

Residualized Gain Scores by Teacher

Score Code	Absolute Freq	Relative Freq (Pct)	Adjusted Freq (Pct)	Cum Freq (Pct)
-10.87	1	3.7	3.8	3.8
-7.26	1	3.7	3.8	7.7
-6.13	1	3.7	3.8	11.5
-4.75	1	3.7	3.8	15.4
-3.70	1	3.7	3.8	19.2
-2.56	1	3.7	3.8	23.1
-2.34	1	3.7	3.8	26.9
-1.94	1	3.7	3.8	30.8
-1.66	1	3.7	3.8	34.6
-1.09	1	3.7	3.8	38.5
-1.04	1	3.7	3.8	42.3
-0.86	1	3.7	3.8	46.2
-0.71	1	3.7	3.8	50.0
-0.43	1	3.7	3.8	53.8
-0.18	1	3.7	3.8	57.7
0.15	1	3.7	3.8	61.5
0.95	1	3.7	3.8	65.4
1.68	1	3.7	3.8	69.2
2.65	1	3.7	3.8	73.1
3.50	1	3.7	3.8	76.9
3.54	1	3.7	3.8	80.8
3.78	1	3.7	3.8	84.6
4.26	1	3.7	3.8	88.5
5.88	1	3.7	3.8	92.3
9.12	1	3.7	3.8	96.2
11.90	1	3.7	3.8	100.0
-9.00	1	3.7	missing	100.0
Total	27	100.0	100.0	

mean	0.072	std err	0.956	median	-0.692
mode	-10.874	std dev	4.875	variance	23.764
kurtosis	0.909	skewness	0.246	range	22.771
minimum	-10.874	maximum	11.898		

valid cases 26 missing cases 1

TABLE 27

Average Stratum Mean Raw Posttest Score by Teacher

Score Code	Absolute Freq	Relative Freq (Pct)	Adjusted Freq (Pct)	Cum Freq (Pct)
41.00	1	3.7	3.7	3.7
42.89	1	3.7	3.7	7.4
46.36	1	3.7	3.7	11.1
46.57	1	3.7	3.7	14.8
46.70	1	3.7	3.7	18.5
47.88	1	3.7	3.7	22.2
47.92	1	3.7	3.7	25.9
48.10	1	3.7	3.7	29.6
49.23	1	3.7	3.7	33.3
50.27	1	3.7	3.7	37.0
50.69	1	3.7	3.7	40.7
51.55	1	3.7	3.7	44.4
52.11	1	3.7	3.7	48.1
52.67	1	3.7	3.7	51.9
53.00	1	3.7	3.7	55.6
53.02	1	3.7	3.7	59.3
53.71	1	3.7	3.7	63.0
53.85	1	3.7	3.7	66.7
54.00	1	3.7	3.7	70.4
55.70	1	3.7	3.7	74.1
56.15	1	3.7	3.7	77.8
56.40	1	3.7	3.7	81.5
57.34	1	3.7	3.7	85.2
58.42	1	3.7	3.7	88.9
59.10	1	3.7	3.7	92.6
62.40	1	3.7	3.7	96.3
66.00	1	3.7	3.7	100.0
Total	27	100.0	100.0	

mean	52.335	std err	1.093	median	52.667
mode	41.000	std dev	5.681	variance	32.277
kurtosis	0.291	skewness	0.278	range	25.000
minimum	41.000	maximum	66.000		

valid cases 27 missing cases 0

Descriptive Statistics

TABLE 28

**Norming Study Total Score Results: Elementary Level
Using Master Teacher Scoring Procedure**

FPMS Score	Freq	Adj Pct	Cum Pct	FPMS Score	Freq	Adj Pct	Cum Pct	FPMS Score	Freq	Adj Pct	Cum Pct
17.	3	1	1	30.	32	6	34	42.	17	3	91
19.	2	0	1	31.	39	7	41	43.	13	2	94
20.	3	1	1	32.	26	5	46	44.	6	1	95
21.	2	0	2	33.	40	7	54	45.	8	1	96
22.	4	1	3	34.	30	6	59	46.	7	1	98
23.	5	1	4	35.	34	6	65	47.	5	1	99
24.	11	2	6	36.	23	4	70	48.	2	0	99
25.	23	4	10	37.	27	5	75	50.	2	0	99
26.	27	5	15	38.	16	3	78	51.	2	0	100
27.	20	4	19	39.	18	3	81	52.	1	0	100
28.	22	4	23	40.	18	3	85	54.	1	0	100
29.	28	5	28	41.	19	4	88				

mean	33.448	std err	0.276	median	33.025
mode	33.000	std dev	6.391	variance	40.846
kurtosis	-0.129	skewness	0.276	range	37.000
minimum	17.000	maximum	54.000		

valid cases 536 missing cases 0

TABLE 29

FPMS Scores By School from MTP Observations Macro Study

17 Schools Having a Minimum of Six Teachers in File

Score Code	Absolute Freq	Relative Freq (Pct)	Adjusted Freq (Pct)	Cum Freq (Pct)
29.	1	2.8	5.9	5.9
32.	1	2.8	5.9	11.8
33.	1	2.8	5.9	17.6
33.	1	2.8	5.9	23.5
34.	1	2.8	5.9	29.4
35.	1	2.8	5.9	35.3
36.	1	2.8	5.9	41.2
36.	1	2.8	5.9	47.1
36.	1	2.8	5.9	52.9
36.	1	2.8	5.9	58.8
37.	1	2.8	5.9	64.7
38.	1	2.8	5.9	70.6
39.	1	2.8	5.9	76.5
40.	1	2.8	5.9	82.4
40.	1	2.8	5.9	88.2
41.	1	2.8	5.9	94.1
42.	1	2.8	5.9	100.0
-9.	19	52.8	missing	100.0
Total	36	100.0	100.0	

mean	36.376	std err	0.858	median	36.352
mode	29.185	std dev	3.539	variance	12.527
kurtosis	-0.372	skewness	-0.205	range	13.229
minimum	29.185	maximum	42.414		

valid cases 17 missing cases 19

TABLE 30

FPMS Scores By School from MTP Observations Fourth to Sixth Grade

17 Schools Having a Minimum of Six Teachers in the Data File

Score Code	Absolute Freq	Relative Freq (Pct)	Adjusted Freq (Pct)	Cum Freq (Pct)
30.	1	2.8	5.9	5.9
33.	1	2.8	5.9	11.8
33.	1	2.8	5.9	17.6
34.	1	2.8	5.9	23.5
34.	1	2.8	5.9	29.4
35.	1	2.8	5.9	35.3
35.	1	2.8	5.9	41.2
37.	1	2.8	5.9	47.1
37.	1	2.8	5.9	52.9
37.	1	2.8	5.9	58.8
38.	1	2.8	5.9	64.7
38.	1	2.8	5.9	70.6
38.	1	2.8	5.9	76.5
39.	2	5.6	11.8	88.2
40.	1	2.8	5.9	94.1
41.	1	2.8	5.9	100.0
-9.	19	52.8	missing	100.0
Total	36	100.0	100.0	

mean	36.256	std err	0.710	median	36.647
mode	39.000	std dev	2.929	variance	8.577
kurtosis	-0.194	skewness	-0.360	range	11.305
minimum	30.029	maximum	41.333		

valid cases 17 missing cases 19

TABLE 31

School Based Residualized Gain Scores for Reading

Final Nine Cases Used in Analysis

RGS Score	Absolute Freq	Relative Freq (Pct)	Adjusted Freq (Pct)	Cum Freq (Pct)
-23.	1	11.1	11.1	11.1
-22.	1	11.1	11.1	22.2
-10.	1	11.1	11.1	33.3
1.	1	11.1	11.1	44.4
5.	1	11.1	11.1	55.6
5.	1	11.1	11.1	66.7
7.	1	11.1	11.1	77.8
21.	1	11.1	11.1	88.9
23.	1	11.1	11.1	100.0
Total	9	100.0	100.0	

Mean	0.860	Std err	5.512	Median	4.556
Mode	-22.810	Std dev	16.536	Variance	273.430
Kurtosis	-0.878	Skewness	-0.259	Range	45.569
Minimum	-22.810	Maximum	22.759		

valid cases 9 missing cases 0

TABLE 32

MTP Scores for All District Elementary Schools

From Two to Seven Teachers Represent Each School

FPMS Score	Freq	Adj Pct	Cum Pct	FPMS Score	Freq	Adj Pct	Cum Pct	FPMS Score	Freq	Adj Pct	Cum Pct
28.80	1	2	2	35.82	1	2	37	39.53	1	2	72
29.18	1	2	4	36.00	2	4	41	39.57	1	2	74
29.50	1	2	6	36.15	1	2	43	39.73	1	2	76
31.50	1	2	7	36.18	1	2	44	40.00	2	4	80
31.55	1	2	9	36.31	1	2	46	40.48	1	2	81
31.83	1	2	11	36.43	1	2	48	40.50	1	2	83
32.91	1	2	13	36.75	1	2	50	40.64	1	2	85
33.00	1	2	15	37.00	1	2	52	40.88	1	2	87
33.25	2	4	19	37.33	1	2	54	41.06	1	2	89
34.00	2	4	22	37.80	1	2	56	42.00	1	2	91
34.42	2	4	26	38.00	2	4	59	42.34	1	2	93
34.76	1	2	28	38.06	1	2	61	42.97	1	2	94
34.82	1	2	30	38.67	1	2	63	43.36	1	2	96
35.00	2	4	33	39.00	3	6	69	44.00	1	2	98
35.60	1	2	35	39.10	1	2	70	44.50	1	2	100

mean	37.018	std err	0.513	median	36.760
mode	39.000	std dev	3.773	variance	14.234
kurtosis	-0.415	skewness	-0.149	range	15.700
minimum	28.800	maximum	44.500		

valid cases 54 missing cases 0

Sample Budget for Conducting
Subject Specific
Predictive Validity Studies

	30 Teachers 4 Observations of each teacher	30 Teachers 8 Observations of each teacher
1. Personnel		
Senior Researcher	12,000.00	12,000.00
Wordprocessor - Secretary	2,500.00	2,500.00
Computer Analyst (Programming, Data Analysis)	10,000.00	10,000.00
Keypunch Operator	300.00	600.00
Field Coordinator	4,800.00	4,800.00
Trainer	300.00	300.00
Trained Observers (6-8,15 obs. each)	1,800.00	3,600.00
Expert Review Team	2,500.00	2,500.00
Subtotal	31,450.00	31,750.00
2. Computer Time	1,000.00	1,500.00
Subtotal	1,000.00	1,500.00
3. Travel and per diem		
Senior Researcher	800.00	800.00
Field Coordinator (8 schools/8 trips)	1,500.00	1,500.00
Trained Observers	1,000.00	2,000.00
Subtotal	3,300.00	4,300.00
4. Miscellaneous Expenses		
Office Supplies (env., opscans, forms)	100.00	100.00
Telephone	50.00	50.00
Xeroxing (pre & post tests, schedules, FYI, training materials)	500.00	500.00
Test scoring/item analysis, etc.	100.00	100.00
Subtotal	750.00	750.00
Total	$36,500.00	$38,300.00

APPENDIX C

Design for a Domain 10.0 Workshop

Session I

Domain 10.0

Time: 8:30 - 8:40 Minutes: 10 Trainer's Notes

Resources:

a. Agenda
b. Participants' materials (domain document, list of indicators, copy of agenda)
c. Paper on overview of Domain 10.0
d. Outline of indicator distinctions

Procedures:

1. Review yesterday briefly. Answer participants questions.

2. Share agenda for the day.

Session II

Domain 10.0

Time: 8:40 - 9:30 Minutes: 40 Trainer's Notes

Procedures:

1. Present overview of Domain 10.0.

2. Walk through list of indicators pointing out distinguishing characteristics and suggest note-taking to the participants.

1. Use paper on overview for presentation.

2. Develop transparencies of major points if desired.

3. Use outline of characteristics for presentation and distribute it if desired.

Session III

Domain 10.0

Time: 9;30 - 10:00	Minutes: 30	Trainer's Notes

Resources:

a. Examples of consultant statements

Procedures:

1. Divide workshop group into small groups to work on written examples of consultant statements. Work for consensus.

2. Do not waste time on items of agreement. Discuss controversial items with the total group.

1. The key for correct answers is a guide for the trainer.

Session IV

Domain 10.0

Time: 10:00 - 10:15	Minutes: 15	Trainer's Notes

Resources:

a. Audiotape
b. Tape Recorder
c. Particpant's copies key to audiotape or transparency

Procedures:

1. For an individual comprehension check, and to develop listening for indicator cues, play the audiotape and have participants write indicator numbers.

2. Give out keys for self check with instructions to take a break and return at 10:30 or provide transparency of answers for self check.

1. Circulate and answer specific questions.

Session V

Domain 10.0

Time: 10:30 - 11:00 Minutes: 30 Trainer's Notes

Resources:

a. Videotape:
   Tarwater(elementary setting)
   Largue(secondary setting)
b. Participants' copies of viewing form
c. VCR and monitor
d. List of indicators

Procedures:

1. Introduce the tape to the group.

2. Show short segments of either tape. Call out indicators or numbers.

3. Rewind tape and play again. This time have participants write indicators or numbers on viewing form. Discuss.

Session VI

Domain 10.0

Time: 11:00 - 12:00 Minutes: 60 Trainer's Notes

Resources:

a. Participants' copies of completed Domain 2.0 and 3.0 instruments
b. Role playing checklist

Procedures:

1. Divide total group into role playing groups.

1. This grouping could be done by putting a number on the handouts and divide numbers 1-10 = Group 1 and 11-20 = Group 2 etc., or use some other grouping plan.

2. Half of the group play conference presenters using data from the instruments. The other half observe the simulation.	2. Workshop leaders should monitor these sessions. It would help to group.
3. The observers critique the presentation.	3. Critique the activity by asking groups to summarize what conferring behaviors were demonstrated adequately, what behaviors were weak, etc. This identifies areas individuals need to practice and emphasize in training.
4. Reverse roles and repeat with the other instrument.	
5. Critique the activity (debrief).	
6. Review the session if workshop terminates at this point. If the next session is scheduled after lunch, delay the review.	

DOMAIN 10.0

Examples of Conferrer's Statements

Using the Concept/Indicator list, match the indicators with the examples. If the example is negative, mark it "n". If the indicator applies to more than one indicator, give both numbers and be prepared to explain your decision.

1. "I like the idea of using student helpers." _____

2. "We have reviewed the data and you indicated that you would be satisfied to work on increasing your use of specific academic praise. Do we agree, then, that this is the problem?" _____

3. "Tell me what you did to check comprehension." _____

4. "Research shows that giving transition signals, as you did in the development of this lesson, shows positive correlation with student comprehension." _____

5. "According to the tally marks in the 2.6-2.7 category on movement smoothness and slowdown, this area of classroom management is marked more often than any of the others." _____

6. "You seem to be saying that you don't agree with the team's assessment of your teaching performance yesterday." _____

7. "Give me some examples of how you might provide for practice." _____

8. "Do we agree that the first area we will work on is maximizing the use of class time for learning? You said that you will work with your peer teacher on improving lesson plans. She will check your plans and report on your progress at the next team meeting which is December 1. I will observe you using the Domain 3.0 formative instrument. Is that right?" _____

9. "I would like to schedule a conference to discuss the lesson with you after the observation. When would it be convenient for you to meet?" _____

10. "You did an effective job of providing work for the students who finished their assignments early. I'm going to borrow your idea of using headphones to provide enrichment activities without distracting the slower workers." ____

11. "I like the way you taught the lesson." ____

12. "Give me some specific examples of the discipline problems you say you are having that seem to be missing from the observation information." ____

13. "Look over the tally marks for a few minutes and then we'll discuss your questions about the indicators." ____

14. "Please review for me your understanding of the areas where you need assistance." ____

15. "I heard you say that you were surprised how many times the data showed you allowed unison response. Is that right?" ____

16. "It is not unusual to feel discouraged when things aren't going smoothly. That is one of the reasons we have a support team to help you identify specific problems and work on solutions together." ____

17. "You don't seem to be having any difficulties in the classroom. I don't think I need to spend any time with you this year. On the other hand that intern from the university doesn't know what to do so she will keep me busy. Do you know her?" ____

18. "You identified talking out without raising you hand as a behavior you would like to work on first. All teachers experience this problem but some learn to handle it effectively. Could you learn from observing others? I could arrange some observations of other teachers for you. We said developing good classroom rules was important for discipline. What ideas do you have about how to improve your rules about talking out?" ____

19. "You asked a lot of high order questions. Research tells us that that is good." _____

20. "From looking at the observation tallies I have identified a problem for us to work on during Cycle I. You are wasting instructional time. I will observe you with formative instrument 3.0 and tell you what to change." _____

Optional Activity

Generate some additional examples in your small group, role play them and see if the others in the group can identify the indicators.

DOMAIN 10.0

Videotape Viewing Form

Directions: As you observe the tape write the behaviors (or their indicator numbers) under the appropriate concept listed below.

Concept 10.1 Development of Rapport

Concept 10.2 Interpersonal Skills

Concept 10.3 Problem Formulation

Concept 10.4 Development of Plans

DOMAIN 10.0

Audiotape Answer Sheet

1. "Is there any evidence of time wasted in this lesson?" ____

2. "Why do you think the students are calling out responses?" ____

3. "Let me see if I have this right. You said you want to work on improving your lesson plans first. You are having trouble pacing the lesson, and also you find that the students don't have the background to discuss the planned lesson without going back over concepts you think they should know. Is that it?" ____

4. "We talked about allowing time for students to think. Do you wish to discuss this area?" ____

5. "The class responded to your more authoritarian behavior today. I noticed that you were able to stop misbehavior by a single word or look." ____

6. "Did I understand you to say that you do not think that the observation data reflect a typical lesson in your classroom?" ____

7. "Yes, you did pause, and research tells us that if we will just pause after a question and give our students time to think, their responses will be better." ____

8. "Yes, you do have tallies for providing feedback, but what exactly do you mean by <u>corrective</u> feedback?" ____

9. "The use of the overhead transparencies to emphasize important points in your lesson was effective." ____

10. "I like to see happy children. You are doing a wonderful job of establishing rapport." ____

DOMAIN 10.0

Role Playing Directions

I. Divide workshop participants into groups of 8:

1 - BLA (building level administrator)
1 - PT (peer teacher)
1 - OE (other educator)
1 - BT (beginning teacher)
4 - Observers

Note: If your workshop is large, increase the number of participants in each group by assigning multiple OE's (other educators) to make a larger support team and/or multiple observers. Be sure to provide enough space for each group to role play without interference.

II. The observers should have a copy of the Role Playing Observer Checklist.

III. The support team should use Domain 3.0 formative instrument as a diagnostic tool in the simulation of a conference with a beginning teacher. The other educator made this observation and the team is now sharing the data. Do not supply other materials.

IV. Use the first 15 minutes (maximum) for role playing.

In the next 10 minutes (maximum) the observers should share their observations.

V. The group then reverses roles - the observers become the role players. Use the Domain 2.0 formative instrument.

VI. After role playing for 15 minutes (maximum) allow 10 minutes (maximum) for sharing the observations.

VII. Call the total group to attention to critique the activity (10 minutes). Encourage the participants to practice the behaviors using their own district materials.

DOMAIN 10.0 ROLE PLAYING CHECKLIST

TALLIES		OBSERVATION NOTES
____	10.1 DEVELOPMENT OF RAPPORT	
____	10.1.1 Arranges Social Space	
____	10.1.2 Promptness	
____	10.1.3 Minimizes Personal Information	
____	10.1.4 Body Behavior	
	10.2 INTERPERSONAL SKILLS	
____	10.2.1 Minimizes Evaluative Expression	
____	10.2.2 Gives Corrective Feedback	
____	10.2.3 Gives Specific Reinforcement	
____	10.2.4 Exercises Flexibility	
____	10.2.5 Maintains Problem Focus	
____	10.2.6 Confines Discussion to Relevant Content	
____	10.2.7 Listens	
____	10.2.8 Makes Specific Clarifying Responses	
____	10.2.9 Induces Teacher Participation	
	10.3 PROBLEM FORMULATION	
____	10.3.1 Examines Observation Data	
____	10.3.2 Elicits Teacher's Concerns about Data	
____	10.3.3 Seeks Clarification of Concerns	
____	10.3.4 Provides Pause Time	
____	10.3.5 Allows Time to Develop the Problem	
____	10.3.6 Elicits Summary of Data Analysis	
____	10.3.7 Restates Summary	
	10.4 DEVELOPMENT OF PLANS	
____	10.4.1 Initiates Planning	
____	10.4.2 Explores Alternatives Collaborately	
____	10.4.3 Summarizes Plans	
____	10.4.4 Confirms Plans	
____	10.4.5 Follow-Up	

LIST OF INDICATORS

DOMAIN 10.0

CONFERRING WITH TEACHERS

10.1 DEVELOPMENT OF RAPPORT

- 10.1.1 Arranges Social Space
- 10.1.2 Promptness
- 10.1.3 Minimizes Personal Information
- 10.1.4 Body Behavior

10.2 INTERPERSONAL SKILLS

- 10.2.1 Minimized Evaluative Expression
- 10.2.2 Gives Corrective Feedback
- 10.2.3 Gives Specific Reinforcement
- 10.2.4 Exercises Flexibility
- 10.2.5 Maintains Problem Focus
- 10.2.6 Confines Discussion to Relevant Content
- 10.2.7 Listens
- 10.2.8 Makes Specific Clarifying Responses
- 10.2.9 Induces Teacher Participation

10.3 PROBLEM FORMULATION

- 10.3.1 Examines Observation Data
- 10.3.2 Elicits Teacher's Concerns About Data
- 10.3.3 Seeks Clarification Of Concerns
- 10.3.4 Provides Pause Time
- 10.3.5 Allows Time To Develop The Problem
- 10.3.6 Elicits Summary Of Data Analysis
- 10.3.7 Restates Summary

10.4 DEVELOPMENT OF PLANS

- 10.4.1 Initiates Planning
- 10.4.2 Explores Alternatives Collaborately
- 10.4.3 Summarizes Plans
- 10.4.4 Confirms Plans
- 10.4.5 Follow-Up